RODRIGO TSAI

APLICAÇÕES DE CÓPULAS EM MODELOS DE RISCOS MÚLTIPLOS DEPENDENTES E EM MODELOS DE MISTURAS DE DISTRIBUIÇÕES

CAMPINAS
2012

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA

RODRIGO TSAI

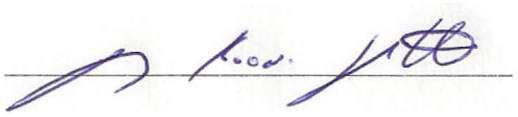
APLICAÇÕES DE CÓPULAS EM MODELOS DE RISCOS MÚLTIPLOS DEPENDENTES E EM MODELOS DE MISTURAS DE DISTRIBUIÇÕES

Orientador: Prof. Dr. Luiz Koodi Hotta

Tese de doutorado apresentada ao Instituto de Matemática, Estatística e Computação Científica da Unicamp para a obtenção do título de Doutor em Estatística.

ESSE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELO ALUNO RODRIGO TSAI, E ORIENTADA PELO PROF. DR. LUIZ KOODI HOTTA.

Assinatura do Orientador:



CAMPINAS
2012

FICHA CATALOGRÁFICA ELABORADA POR
ANA REGINA MACHADO - CRB8/5467
BIBLIOTECA DO INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E
COMPUTAÇÃO CIENTÍFICA - UNICAMP

Tsai, Rodrigo, 1974-
T782a Aplicações de cópulas em modelos de riscos múltiplos dependentes e em modelos de misturas de distribuições / Rodrigo Tsai. – Campinas, SP : [s.n.], 2012.

Orientador: Luiz Koodi Hotta.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Análise de sobrevivência (Biometria). 2. Variáveis latentes. 3. Riscos competitivos. 4. Distribuição (Probabilidades). I. Hotta, Luiz Koodi,1952-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações para Biblioteca Digital

Título em inglês: Applications of copula to polyhazard models with dependence and mixture models
Palavras-chave em inglês:
Survival analysis (Biometry)
Latent variables
Competing risks
Distribution (Probability theory)
Área de concentração: Estatística
Titulação: Doutor em Estatística
Banca examinadora:
Luiz Koodi Hotta [Orientador]
Hildete Prisco Pinheiro
Mário de Castro Andrade Filho
Francisco Louzada Neto
Antonio Carlos Pedroso de Lima
Data de defesa: 30-11-2012
Programa de Pós-Graduação: Estatística

Tese de Doutorado defendida em 30 de novembro de 2012 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). LUIZ KOODI HOTTA

Prof(a). Dr(a). HILDETE PRISCO PINHEIRO

Prof(a). Dr(a). MÁRIO DE CASTRO ANDRADE FILHO

Prof(a). Dr(a). FRANCISCO LOUZADA NETO

Prof(a). Dr(a). ANTONIO CARLOS PEDROSO DE LIMA

Agradecimentos

Ao meu orientador Prof. Luiz Koodi Hotta pela paciência, dedicação, orientação e amizade.

Aos professores da comissão examinadora e qualificação por sua disponibilidade e contribuição.

Aos professores do Departamento de Estatística da Unicamp pelos ensinamentos e amizade.

À Capes pela bolsa concedida em parte do programa, aos funcionários do IMECC e ao Laboratório EPIFISMA.

À minha família e aos amigos pela compreensão, apoio e incentivo.

A Deus por todas essas pessoas e por tudo que pude realizar.

Resumo

TSAI, R. **Aplicações de Cópulas em Modelos de Riscos Múltiplos Dependentes e em Modelos de Misturas de Distribuições**. 2012. 126 f. Tese (Doutorado) - Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Campinas, 2012.

Nesse trabalho discutimos aplicações de cópulas a modelos de riscos múltiplos com dependência e modelos de misturas de distribuições. Numa primeira parte analisamos a inclusão de dependência entre os fatores de risco do modelo de riscos múltiplos. Os modelos de riscos múltiplos são uma família de modelos flexíveis para representar dados de tempos de vida. Suas maiores vantagens sobre os modelos de risco simples incluem a habilidade de representar funções de taxa de falha com formas não usuais e a facilidade de incluir covariáveis. O objetivo principal dessa parte é modelar a dependência existente entre as causas latentes de falha do modelo de riscos múltiplos por meio de funções de cópulas. A escolha da função de cópulas bem como das funções de distribuição dos tempos latentes de falha resultam numa classe flexível de distribuições de sobrevivência que é capaz de representar funções de taxa de falha de formas multimodais, forma de banheira e contendo efeitos locais dados pela concorrência dos riscos. A identificação e estimação do modelo proposto também são discutidas. Ao eliminar a restrição de suporte positivo para as variáveis latentes, o método pode ser utilizado para gerar uma família rica de distribuições univariadas contendo assimetrias e múltiplas modas. Na segunda parte propomos um modelo de mistura de distribuições generalizado utilizando cópulas. O parâmetro da cópula é útil para definir formas de assimetria e ponderar com maior ou menor peso determinadas regiões do suporte das distribuições componentes para compor a mistura. pesos das distribuições componentes variam no suporte da distribuição e não são restritos à soma unitária. A modelagem resultante acrescenta uma maior flexibilidade aos modelos de misturas na representação de dados com densidades de várias formas multimodais e assimétricas. O modelo tem como casos particulares o modelo de mistura tradicional, o modelo de riscos múltiplos e o modelo de fração de cura. Os modelos são aplicados a dados simulados e reais da literatura. Foram utilizados os métodos de estimação de máxima verossimilhança e os critérios de ajuste de Akaike e Bayesiano para a seleção dos modelos. Os modelos representaram bem os conjuntos de dados analisados em comparação com metodologias propostas na literatura.

Palavras-chave: cópulas, riscos competitivos, modelos de riscos múltiplos, modelos de misturas.

Abstract

TSAI, R. **Applications of Copula to Polyhazard Models with Dependence and Mixture Models**. 2012. 126 f. Tese (Doutorado) - Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Campinas, 2012.

In this work, we discuss the application of copula to polyhazard and mixture models. First we analyse the inclusion of dependence among failure causes in the polyhazard models. The polyhazard models constitute a family of flexible models to represent lifetime data. Their main advantages over single hazard models include the ability to represent hazard rate functions with unusual shapes and the ease of including covariates. The main purpose in this first part is to model the dependence that exists among the latent causes of failure in the polyhazard model by copula functions. The choice of the copula function as well as the latent failure distributions produces a flexible class of survival distributions that is able to model hazard functions with unusual shapes such as bathtub or multimodal curves, while also modelling local effects given by the competing risks. The model identification and estimation are also discussed. Dropping the restriction of positive support for the latent variables, the method can be used to generate a rich family of univariate distributions with asymmetries and multiple modes. In the second part a generalized mixture model using copula functions is proposed. To assemble the mixture model, the parameter of the copula function is used to define asymmetry shapes and to attribute more or less weight to chosen regions of the component distributions. The weights of the component distributions vary on the support of the distribution and are not restricted to the unitary sum. The resulting model increases the flexibility of the mixture models to represent data with densities with several multimodal and asymmetric shapes. Special cases of the model are the traditional mixture models, the polyhazard model, and the cure fraction model. Simulated and empirical data from the literature are analysed by the proposed models. The estimation was done by maximum likelihood methods and the selection of the models used the Akaike and Bayesian criteria. The proposed models exhibited very good fit to the data sets in comparison to other methodologies presented in the literature.

Keywords: copula, competing risks, polyhazard models, mixture models.

Sumário

Lista de Abreviaturas

SJC	Cópula de Joe-Clayton simetrizada (*Symmetrized Joe-Clayton copula*)
GBG	Modelo gerado de distribuição beta generalizado (*Generalized beta-generated model*)
BG	Modelo gerado de distribuição beta (*Beta-generated model*)
PHD	Modelo de riscos múltiplos com dependência (*Polyhazard model with dependence*)
GBW	Modelo beta Weibull generalizado (*Generalized beta Weibull model*)
GBN	Modelo beta normal generalizado (*Generalized beta normal model*)
GBLN	Modelo beta log-normal generalizado (*Generalized beta log-normal model*)
GBGam	Modelo beta gamma generalizado (*Generalized beta gamma model*)
GBGum	Modelo beta Gumbel generalizado (*Generalized beta Gumbel model*)
GBTSk	Modelo beta t-assimétrica generalizado (*Generalized beta t-skewed model*)
GBSLa	Modelo beta Laplace escalada generalizado (*Generalized beta Scaled Laplace model*)
BMW	Modelo beta modificada Weibull (*Beta-modified Weibull model*)
BGGam	Modelo beta gamma generalizada (*Beta-generalized gamma model*)
MTrad	Modelo de mistura tradicional
MGC	Modelo de mistura generalizado por cópulas
MGC-CS	Modelo de mistura generalizado por cópulas usando conjuntos simples
MGC-CS-O	Modelo de mistura de valores de regiões opostas das distribuições usando conjuntos simples
MGC-CS-CED	Modelo de mistura de valores de regiões à esquerda ou à direita de uma das componentes usando conjuntos simples
MGC-CS-EA	Modelo de mistura de efeito de assimetria usando conjuntos simples
MGC-CR	Modelo de mistura generalizado por cópulas usando conjuntos de reflexão
MGC-CR-EE	Modelo de mistura de valores de regiões à esquerda de distribuições usando conjuntos de reflexão
MGC-CR-ED	Modelo de mistura de valores de regiões à esquerda e à direita de distribuições usando conjuntos de reflexão
MGC-CR-DD	Modelo de mistura de valores de regiões à direita de distribuições usando conjuntos de reflexão

MGC-CR-CAE	Modelo de mistura de valores das caudas de distribuições com assimetria à esquerda usando conjuntos de reflexão
MGC-CR-CAD	Modelo de mistura de valores das caudas de distribuições com assimetria à direita usando conjuntos de reflexão
MGC-3CS-CC	Modelo de mistura de valores das caudas e centro das distribuições usando três conjuntos simples
MGC-3CS-CC-S	Modelo de mistura de valores das caudas e centro das distribuições usando conjuntos simples com tomada simétrica dos quantis

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Nessa tese é utilizada a abordagem de cópulas para propor dois modelos generalizados. Esses modelos são o modelo de riscos múltiplos, ou *polyhazard*, com dependência, e o modelo de mistura de distribuições generalizado.

O modelo de riscos múltiplos é desenvolvido na literatura de análise de sobrevivência e confiabilidade para representar tempos de vida em situações que há causas de falha latentes atuando sobre os indivíduos em observação, chamadas de riscos competitivos, e que a causa da falha ocorrida para cada individuo é desconhecida. Há vários exemplos de modelos de riscos múltiplos independentes que aparecem na literatura desde, por exemplo, Kalbfleisch e Prentice (1980), que propuseram o modelo poli-log-logístico para riscos competitivos log-logísticos e Berger e Sun (1993), que propuseram o modelo poli-Weibull para riscos competitivos Weibull e sua estimação por análise Bayesiana usando o amostrador de Gibbs.

Uma primeira contribuição metododológica desse trabalho é a proposta do modelo de riscos múltiplos com dependência. Nesse trabalho a denominação de riscos é utilizada para significar variáveis aleatórias latentes que representam tempos. O modelo de riscos múltiplos com dependência é desenvolvido a partir do modelo de riscos múltiplos independentes, que tem formulação baseada em tempos latentes de falha. O método é então baseado na utilização de funções de cópulas para modelar o comportamento conjunto das variáveis latentes. Resulta dessa metodologia uma família flexível para ajustar dados de tempos de vida, representando funções taxa de falha não usuais com forma de banheira, multimodal e contendo efeitos locais, dados pelos riscos competitivos.

Em seguida, essa metodologia proposta de riscos múltiplos dependentes, destinada à análise de tempos de vida, é estendida a variáveis latentes que podem assumir valores em toda a reta real. Dessa extensão do método resulta uma família de modelos univariados com boas qualidades para representar conjuntos de dados sem a restrição a valores positivos.

A outra metodologia proposta nessa tese é o modelo de misturas generalizado utilizando cópulas. Na literatura, os modelos de misturas finitas estão presentes em várias áreas da estatística como análise de agrupamentos e estruturas latentes, análise discriminante, análise de sobrevivência e também são úteis em inferência e análise de dados em geral, providenciando modelos descritivos para distribuições dos dados.

A modelagem nessa proposta resulta de uma transformação de variáveis aleatórias distribuídas segundo uma cópula. Nessa transformação são utilizadas funções de distribuição que participam como distribuições componentes da mistura. A distribuição dessa variável transformada equivale à mistura dessas distribuições componentes com pesos que variam no suporte da distribuição e não são restritos à soma unitária. Nessa formulação, os parâmetros da função de cópula e da mistura proposta são úteis para definir as formas de assimetria e peso de cada componente transformada no modelo. O modelo proposto é considerado um modelo de misturas generalizado.

Como resultados dessa metodologia, o modelo generalizado inclui outros modelos da literatura como

casos particulares, que são o modelo de riscos múltiplos com dependência, o modelo de fração de cura, especificado por mistura de distribuições, e o modelo de misturas tradicional, que atribui pesos constantes às distribuições componentes. Resultam também dessa metodologia famílias de modelos derivadas dessa proposta com boas características de ajuste a dados, capazes de representar multimodalidade e assimetria dos dados. Nessas famílias, os parâmetros da cópula e da mistura são utilizados para definir as características de assimetria e ponderar com maior ou menor peso determinadas regiões das distribuições componentes para compor o modelo de mistura.

Para concluir, a tese está organizada da seguinte forma. No Capítulo 2 são apresentados alguns conceitos básicos utilizados na tese sobre cópulas, análise de sobrevivência e modelos de misturas. O Capítulo 3 apresenta a primeira metodologia proposta nessa tese. O modelo de riscos múltiplos com dependência é definido e ilustrado, são discutidas sua identificação e estimação, e por fim o modelo é aplicado a dados simulados e a um conjunto de dados de duração de desemprego de parte da força de trabalho da Alemanha. Em seguida, a extensão do modelo a variáveis assumindo valores nos reais é proposta e analisada em comparação com outros modelos da literatura, em sua capacidade de representar funções densidade e taxa de falha. O modelo estendido também é ajustado a conjuntos de dados reais da literatura em comparação com outros métodos. O Capítulo 4 apresenta a segunda metodologia proposta nessa tese, do modelo de misturas genralizado utilizando cópulas. Inicialmente, os modelos de mistura tradicional, fração de cura e de riscos múltiplos do Capítulo 3 são analisados como casos particulares do modelo de misturas generalizado. Em seguida, algumas famílias do modelo de misturas generalizado são propostas e ilustradas, e sua identificação e estimação são discutidas. O ajuste dessas famílias a dados simulados e empíricos da literatura de modelos de misturas também é analisado em comparação com outras metodologias. Finalmente, no Capítulo 5 são discutidas algumas conclusões obtidas nesse trabalho e sugestões de pesquisas futuras. Os Apêndices A e B contêm os artigos desenvolvidos pelo candidato e orientador sobre o modelo de riscos múltiplos com causas de falha dependendentes discutido no Capítulo 3.

Capítulo 2

Conceitos Básicos

Nesse capítulo são apresentados alguns conceitos fundamentais relacionados aos métodos propostos nessa tese. Na Seção 2.1 são apresentadas algumas definições e características de funções de cópulas. Essas funções são utilizadas no modelo de riscos múltiplos com dependência do Capítulo 3 e no modelo de misturas generalizado do Capítulo 4. Na Seção 2.2 são apresentados os conceitos básicos de análise de sobrevivência e modelagem de riscos competitivos. Esses temas são relacionados ao modelo de riscos múltiplos com dependência do Capítulo 3. Finalmente, a Seção 2.3 apresenta os modelos de misturas finitas. Esses modelos em sua forma univariada são casos particulares do modelo de misturas generalizado do Capítulo 4.

2.1 Cópulas

Cópulas são funções que conectam distribuições multivariadas a suas marginais unidimensionais. Métodos baseados em cópulas têm sido amplamente utilizados na literatura. O interesse com essas funções surge de várias perspectivas. Primeiro, em situações em que há maior disponibilidade de informações sobre o comportamento marginal de variáveis em estudo que informações sobre o comportamento conjunto dessas variáveis. A abordagem de cópulas é um método útil de derivar distribuições conjuntas a partir de distribuições marginais, principalmente quando as variáveis não são normais. Também, no contexto bivariado, cópulas podem ser usadas para obter medidas não paramétricas de dependência para pares de variáveis aleatórias. Em situações em que são necessárias medidas de dependência mais gerais do que correlação linear, cópulas são bastante úteis para desenvolver conceitos e medidas adequadas. Em terceiro lugar, cópulas permitem extensões e generalizações de abordagens de modelagem de distribuições conjuntas e dependência da literatura. Joe (1997), Cherubini *et al.* (2004) e Nelsen (2006) são dedicados à teoria de cópulas.

Para a definição das funções de cópulas é necessário definir função fundada em seu domínio *(grounded)*, o volume de uma função bivariada sobre um retângulo e a propriedade de crescimento bidimensional *(2-increasing)* de funções bivariadas. O conjunto dos números reais estendido é denotado por $\bar{R}$.

Definição 1 *Sejam S_1 e S_2 conjuntos não vazios de $\bar{R}$. Uma função H definida em $S_1 \times S_2$ e tomando valores em $\mathcal{R}$ é dita fundada em seu domínio se para a_1, o menor elemento em S_1, e a_2, o menor elemento em S_2, tivermos $H(a_1, v) = 0 = H(u, a_2)$ para quaisquer u em S_1 e v em S_2.*

Definição 2 *Sejam S_1 e S_2 dois conjuntos não vazios de $\bar{R}$, H uma função real bivariada tal que o domínio de H é $Dom(H) = S_1 \times S_2$. Seja $B = [x_1, x_2] \times [y_1, y_2]$ um retângulo cujos vértices pertencem a $Dom(H)$. Então, o volume-H de B é dado por*

$$V_H(B) = H(x_2, y_2) - H(x_2, y_1) - H(x_1, y_2) + H(x_1, y_1). \tag{2.1}$$

Definição 3 *Uma função bivariada H tem crescimento bidimensional se $V_H(B) \geq 0$ para todos os retângulos B cujos vértices pertencem ao domínio de H.*

Dessa forma, uma função com crescimento bidimensional atribui valor não negativo para qualquer retângulo em seu domínio. A seguir são definidas as funções de subcópulas e de cópulas, como subcópulas definidas em $I^2 = [0,1]^2$.

Definição 4 *Uma subcópula bidimensional é uma função $\check{C}$ com as seguintes propriedades:*

1. *$Dom(\check{C}) = S_1 \times S_2$, em que S_1 e S_2 são subconjuntos de $I = [0,1]$ contendo 0 e 1;*
2. *$\check{C}$ é fundada em seu domínio e tem crescimento bidimensional;*
3. *Para todo u em S_1 e v em S_2*

$$\check{C}(u,1) = u \quad e \quad \check{C}(1,v) = v. \tag{2.2}$$

Note que para todo (u,v) em $Dom(\check{C})$, $0 \leq \check{C}(u,v) \leq 1$, de forma que a imagem de $\check{C}$ também é um subconjunto de I.

Definição 5 *Uma cópula bidimensional é uma subcópula bidimensional cujo domínio é I^2.*

Dessa forma, uma cópula é uma função C de I^2 em I com as propriedades:

1. Para todos u, v em I

$$C(u,0) = C(0,v) = 0, \tag{2.3}$$

e

$$C(u,1) = u \quad e \quad C(1,v) = v. \tag{2.4}$$

2. Para todos u_1, u_2, v_1, v_2 em I tais que $u_1 < u_2$ e $v_1 < v_2$,

$$V_C([u_1,u_2] \times [v_1,v_2]) = C(u_2,v_2) - C(u_1,v_2) - C(u_2,v_1) + C(u_1,v_1) \geq 0. \tag{2.5}$$

Também temos que para toda cópula C e todo (u,v) em I^2,

$$\max(u+v-1,0) \leq C(u,v) \leq \min(u,v), \tag{2.6}$$

em que os limites $W(u,v) = \max(u+v-1,0)$ e $M(u,v) = \min(u,v)$ são chamados limites de Fréchet-Hoeffding e também são cópulas. O conhecimento desses limites é importante para selecionar uma cópula apropriada. É desejável que uma cópula seja capaz de representar todo o espaço entre esses limites inferior e superior, embora isso não ocorra para muitas cópulas. Uma outra cópula importante é a cópula produto, $\Pi(u,v) = C(u,v) = uv$.

Definida a função de cópulas, a seguir é apresentado o teorema de Sklar.

Teorema 2.1.1 *Seja H uma função de distribuição conjunta com marginais F e G. Então, existe uma cópula C tal que para quaisquer x,y em $\bar{R}$,*

$$H(x,y) = C(F(x),G(y)). \tag{2.7}$$

Se F e G são contínuas, então C é única; se não são contínuas, então C é unicamente determinada em $Im(F) \times Im(G)$. Reciprocamente, se C é uma cópula e F e G são funções de distribuição, então a função H definida por (2.7) é uma função de distribuição conjunta com marginais F e G.

Logo, para uma distribuição conjunta H de marginais contínuas F e G, existe uma cópula C que satisfaz (2.7) e ela é única. Também, para as distribuições marginais F e G, $C(F(x), G(y))$ é uma função de distribuição conjunta.

Para considerar variáveis aleatórias nessa abordagem é útil denotar C_{XY} para a cópula C tal que $H(x, y) = C(F(x), G(y))$, em que as variáveis aleatórias X e Y têm distribuições marginais F e G, respectivamente, e distribuição conjunta H. Assim C_{XY} é chamada de cópula de X e Y. O teorema a seguir mostra que a cópula produto $\Pi(u, v) = uv$ caracteriza variáveis aleatórias independentes quando as funções de distribuição são contínuas. Sua prova é dada pelo Teorema 2.1.1 e pelo conhecimento que X e Y são variáveis aleatórias independentes se, e somente se, $H(x, y) = F(x)G(y)$.

Teorema 2.1.2 *Sejam X e Y variáveis aleatórias contínuas. Então X e Y são independentes se, e somente se, $C_{XY} = \Pi$.*

Uma outra propriedade importante de cópulas é de sua invariância com respeito a transformações estritamente monótonas.

Teorema 2.1.3 *Sejam X e Y variáveis aleatórias contínuas com cópula C_{XY}. Se α e β são estritamente crescentes na imagem de X e Y, respectivamente, então $C_{\alpha(X)\beta(Y)} = C_{XY}$. Logo, C_{XY} é invariante com respeito a transformações estritamente crescentes em X e Y.*

Prova. Sejam F_1, G_1, F_2 e G_2 as funções de distribuição de X, Y, $\alpha(X)$ e $\beta(Y)$, respectivamente. Devido a α e β serem estritamente crescentes, $F_2(x) = P[\alpha(X) \leq x] = P[X \leq \alpha^{-1}(x)] = F_1(\alpha^{-1}(x))$. Semelhantemente, $G_2(y) = G_1(\beta^{-1}(y))$. Então para quaisquer x e y em R,

$$\begin{aligned} C_{\alpha(X)\beta(Y)}(F_2(x), G_2(y)) &= P[\alpha(X) \leq x, \beta(Y) \leq y] \\ &= P[X \leq \alpha^{-1}(x), Y \leq \beta^{-1}(y)] \\ &= C_{XY}(F_1(\alpha^{-1}(x)), G_1(\beta^{-1}(y))) \\ &= C_{XY}(F_2(x), G_2(y)) \quad \blacksquare \end{aligned}$$

Quando X e Y são contínuas, Y é quase certamente uma função crescente de X se, e somente se, a cópula de X e Y é M, e Y é quase certamente uma função decrescente de X se, e somente se, a cópula de X e Y é W. Variáveis com cópula M são chamadas de comonotônicas e variáveis com cópula W são chamadas de contramonotônicas.

Quando pelo menos uma das funções α e β é estritamente decrescente, resulta que a cópula das transformações $\alpha(X)$ e $\beta(Y)$ são transformações simples de X e Y.

Teorema 2.1.4 *Sejam X e Y variáveis aleatórias contínuas com cópula C_{XY}. Sejam α e β funções estritamente monótonas nas imagens de X e Y, respectivamente.*

1. *Se α é estritamente crescente e β é estritamente decrescente, então*

$$C_{\alpha(X)\beta(Y)}(u, v) = u - C_{XY}(u, 1 - v).$$

2. *Se α é estritamente decrescente e β é estritamente crescente, então*

$$C_{\alpha(X)\beta(Y)}(u, v) = v - C_{XY}(1 - u, v).$$

3. *Se α é estritamente decrescente e β é estritamente crescente, então*

$$C_{\alpha(X)\beta(Y)}(u, v) = u + v - 1 + C_{XY}(1 - u, 1 - v).$$

Para a função de sobrevivência conjunta de X e Y, $\bar{H}(x,y) = P[X > x, Y > y]$, e suas funções de sobrevivência marginais $\bar{F}$ e $\bar{G}$, respectivamente, uma pergunta natural é se existe um relacionamento entre essas funções que seja análogo ao estabelecido pelo teorema de Sklar para funções de distribuição marginais e conjunta. Para C a cópula de X e Y temos

$$\begin{aligned}\bar{H}(x,y) &= 1 - F(x) - G(y) + H(x,y) \\ &= 1 - F(x) - G(y) + C(F(x), G(y)) \\ &= \bar{F}(x) + \bar{G}(y) - 1 + C(1 - \bar{F}(x), 1 - \bar{G}(y)). \end{aligned} \tag{2.8}$$

Se definimos $\widehat{C}(u,v) = u + v - 1 + C(1-u, 1-v)$, então temos que

$$\bar{H}(x,y) = \widehat{C}(\bar{F}(x), \bar{G}(y)), \tag{2.9}$$

em que $\widehat{C}$ pode ser verificada ser uma cópula, chamada de cópula de sobrevivência de X e Y. $\widehat{C}$ acopla a função de sobrevivência conjunta às suas marginais. É importante notar a diferença entre a cópula de sobrevivência $\widehat{C}$ e a função de sobrevivência conjunta $\bar{C}$ de duas variáveis aleatórias uniformes em $(0,1)$, em que $\bar{C} = P[U > u, V > v] = 1 - u - v + C(u,v) = C(1-u, 1-v)$.

A Tabela 2.1 mostra algumas funções de cópulas uniparamétricas principais. Essas cópulas são utilizadas no modelo de riscos múltiplos com dependência do Capítulo 3 e no modelo de misturas generalizado do Capítulo 4.

2.2 Análise de Sobrevivência

Consideramos nessa seção alguns conceitos básicos e métodos utilizados em análise de sobrevivência. Dois livros texto nesse assunto são Klein e Moeschberger (2003) e Hosmer *et al.* (2008). A seguir apresentamos as funções principais usadas na modelagem de dados de sobrevivência. Em seguida é apresentada uma visão geral dos métodos paramétricos e semiparamétricos mais comuns em análise de tempos de vida e a modelagem de riscos competitivos.

Suponha que X é o tempo decorrido até a realização de um evento de interesse. Esse evento pode ser a morte do indivíduo em observação, o aparecimento de um tumor, a quebra de um equipamento, cessação de amamentação etc. Na literatura de análise de sobrevivência esse evento normalmente é chamado de falha, embora não seja o objetivo caracterizar uma situação negativa. Também, o tempo de observação até a ocorrência do evento é frequentemente chamado de tempo de falha. Consideramos que X é uma variável aleatória não negativa de uma população homogênea. Na análise de sobrevivência, quatro funções caracterizam a distribuição de X, que são, a função de sobrevivência, que é a probabilidade de um indivíduo sobreviver a um tempo x; a função taxa de falha ou função de risco, que é a taxa instantânea de falha no tempo x dado que o indivíduo está vivo no tempo x; a função de densidade de probabilidade, que é a probabilidade incondicional da ocorrência do evento no tempo x; e o tempo de vida médio residual em x, que é o tempo médio até o evento de interesse dado que o evento não ocorreu até x. Se conhecemos uma dessas funções, então as outras três funções podem ser unicamente determinadas. Na prática, essas quatro funções, em conjunto com outra quantidade importante e também equivalente às demais, a função de taxa de falha acumulada, são utilizadas para ilustrar aspectos diferentes da distribuição de X. Ainda uma outra função, a taxa de falha de causa específica, é utilizada no contexto de riscos competitivos, em que é considerado que um número de causas de falha atuam no indivíduo em observação de forma que concorrem para ser a causa da falha. Essa função é a taxa de falha devida à i-ésima causa de falha em sujeitos para os quais ainda não foi observada a ocorrência do evento. Essas funções são discutidas a seguir.

2.2.1 A Função de Sobrevivência

A quantidade básica utilizada para descrever os fenômenos de tempo até o evento de interesse é a função de sobrevivência, que é a probabilidade de um indivíduo sobreviver além do tempo x. Ela é definida como

$$S(x) = P(X > x). \tag{2.10}$$

No contexto de falhas de produtos ou equipamentos manufaturados, $S(x)$ é chamada função de confiabilidade. Se X é uma variável aleatória contínua, então $S(x)$ é uma função contínua estritamente decrescente no seu domínio. A função de sobrevivência é o complemento da função de distribuição acumulada, isto é, $S(x) = 1 - F(x)$, em que $F(x) = P(X \leq x)$. Também, a função de sobrevivência é a integral da função densidade de probabilidade, $f(x)$, que é

$$S(x) = \int_0^x f(t)dt, \tag{2.11}$$

do que segue que

$$f(x) = -\frac{dS(x)}{dx}. \tag{2.12}$$

Note que $f(x)dx$ pode ser vista como uma probabilidade aproximada que o evento irá ocorrer no tempo dx e que $f(x)$ é uma função não negativa tal que a área sob $f(x)$ é igual a um.

Muitos tipos de curvas de sobrevivência podem ser discutidos, mas todos têm as mesmas propriedades básicas. No caso geral, em que X é variável aleatória contínua ou discreta, essas funções são monótonas, não crescentes, iguais a um no tempo zero e valem zero quando o tempo tende ao infinito. Sua taxa de decrescimento varia, obviamente, de acordo com o risco de ocorrência do evento no tempo x, mas é difícil determinar o comportamento do padrão de falha simplesmente observando a curva de sobrevivência. Apesar disso, essa quantidade é uma descrição bastante utilizada na literatura de sobrevivência aplicada e é muito útil para comparar dois ou mais padrões de mortalidade.

Quando X é uma variável aleatória discreta, técnicas diferentes são requeridas. Essas variáveis aparecem em análise de sobrevivência devido a arredondamentos, agrupamento de tempos de falha em intervalos, ou quando os tempos de vida se referem a um número de unidades integradas. Suponha que X possa tomar valores x_j, $j = 1, 2, ...$ com função de massa de probabilidade $p(x_j) = P(X = x_j)$, $j = 1, 2, ...$, em que $x_1 < x_2 < ...$. A função de sobrevivência para variáveis aleatórias discretas é uma função escada não crescente dada por

$$S(x) = P(X > x) = \sum_{x_j < x} p(x_j). \tag{2.13}$$

2.2.2 A Função de Taxa de Falha

Uma quantidade fundamental em análise de sobrevivência é a função de taxa de falha. Essa função é também conhecida como a taxa de falha condicional em confiabilidade, força de mortalidade em demografia, função intensidade em processos estocásticos, taxa de falha de idade específica em epidemiologia, razão inversa de Mill em economia ou simplesmente taxa de risco. A taxa de risco é definida por

$$h(x) = \lim_{\Delta x \to 0} \frac{P[x \leq X < x + \Delta x | X \geq x]}{\Delta x}. \tag{2.14}$$

Se X é uma variável aleatória contínua, então

$$h(x) = \frac{f(x)}{S(x)} = -\frac{d\ \log[S(x)]}{dx}. \tag{2.15}$$

Uma quantidade relacionada a $h(x)$ é a função de taxa de falha acumulada $H(x)$, definida por

$$H(x) = \int_0^x h(u)du = -ln[S(x)]. \tag{2.16}$$

Do que temos, para tempos de vida contínuos,

$$S(x) = \exp[-H(x)] = \exp\left[-\int_0^x h(u)du\right]. \tag{2.17}$$

De (2.14), temos que $h(x)\Delta x$ pode ser vista como a probabilidade aproximada de, para um indivíduo de idade x, se observar o evento no próximo intervalo Δx. Essa função é particularmente útil em determinar as distribuições apropriadas de falha utilizando informação qualitativa sobre o mecanismo de falha e também para descrever a forma pela qual a probabilidade de ocorrência do evento varia com o tempo. Existem muitas formas para a função taxa de falha, sendo a única restrição que $h(x)$ seja não negativa, i.e., $h(x) > 0$. Alguns tipos básicos de taxas de falha descrevem situações em que a taxa é crescente, decrescente, em forma de banheira, em forma unimodal, havendo outras formas com características que descrevem o mecanismo particular de falha.

Modelos com taxa de falha crescente ocorrem em situações que existe o envelhecimento natural. Funções de taxa de falha decrescentes são muito menos comuns, e são usadas quando há chance de falha em tempos curtos, como em certos tipos de eletrônicos ou pacientes sob situações de transplantes de órgãos. Funções de taxas de falha em forma de banheira são observados com freqüência e ocorrem em populações que são acompanhadas desde o nascimento dos indivíduos. Similarmente, alguns equipamentos manufaturados podem passar por falhas prematuras devidas a partes contendo defeitos, seguidos por uma taxa de falha constante que, aumenta em estágios mais longos. A maioria dos dados de mortalidade de populações segue esse tipo de função de taxa de falha em que, durante estágios iniciais as mortes resultam principalmente de doenças da infância, depois dos quais a taxa de morte estabiliza, seguido por um crescimento devido ao processo natural de envelhecimento. Finalmente, se a taxa de falha é inicialmente crescente e depois seguida de um declínio, então a taxa de falha é chamada ter forma unimodal. Esse tipo de taxa de falha é frequentemente utilizado na modelagem de sobrevivência após cirurgias bem sucedidas em que, inicialmente, existe um risco crescente devido a infecções, hemorragias ou outras complicações que ocorrem no período após o procedimento, seguido por um declínio no risco de acordo com a recuperação do paciente.

Quando X é uma variável aleatória discreta, a função de taxa de falha é dada por

$$h(x_j) = P(X = x_j | X \geq x_j) = \frac{p(x_j)}{S(x_{j-1})}, j = 1, 2, ..., \tag{2.18}$$

em que $S(x_0) = 1$. Dado que $p(x_j) = S(x_{j-1}) - S(x_j)$, segue que $h(x_j) = 1 - S(x_j)/S(x_{j-1}), j = 1, 2, ...$ Note que para uma variável aleatória discreta a taxa de falha é zero exceto nos pontos em que a falha pode ocorrer. A função de sobrevivência pode ser escrita em função de probabilidades condicionais

$$S(x) = \prod_{x \geq x_j} \frac{S(x_j)}{S(x_{j-1})}, \tag{2.19}$$

do que segue que a função de sobrevivência pode ser relacionada à taxa de falha por

$$S(x) = \prod_{x \geq x_j} [1 - h(x_j)]. \tag{2.20}$$

2.2.3 A Função Tempo de Vida Residual Médio

A quarta função básica de interesse em análise de sobrevivência é o tempo de vida residual médio no instante x. Para indivíduos de idade x, essa função mede o tempo de vida restante esperado. Ele é definido como $mrl(x) = E[X - x|X > x]$. Pode ser mostrado que o tempo de vida residual médio é a área sob a curva de sobrevivência à direita de x dividida por $S(x)$. Note que o tempo médio de vida $\mu = mrl(0)$, é a área sob toda a curva de sobrevivência. Para uma variável aleatória contínua,

$$mrl(x) = \frac{\int_x^\infty (t-x) f(t) dt}{S(x)} = \frac{\int_x^\infty S(t) dt}{S(x)} \tag{2.21}$$

e

$$\mu = E[X] = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt. \tag{2.22}$$

Também a variância de X é relacionada à função de sobrevivência por

$$Var(X) = 2 \int_0^\infty t S(t) dt - \left[\int_0^\infty S(t) dt \right]^2. \tag{2.23}$$

O quantil de probabilidade p, ou percentil de probabilidade $100p\%$ da distribuição de X é o menor x_p tal que

$$S(x_p) \geq 1 - p, \; i.e., x_p = \inf[t; S(t) \leq 1 - p]. \tag{2.24}$$

Se X é uma variável aleatória contínua, então o quantil de probabilidade p é encontrado resolvendo ae equação $S(x_p) = 1 - p$. O tempo de vida mediano é o percentil de probabilidade $50p\%$, $x_{0,5}$, da distribuição de X. Segue que o tempo mediano para uma variável aleatória contínua é o valor $x_{0,5}$ tal que $S(x_{0,5}) = 0,5$.

2.2.4 Modelos Paramétricos

Alguns modelos paramétricos frequentemente utilizados em análise de sobrevivência são os modelos exponencial, Weibull, log-normal, log-logística e gama. Mais adiante nessa tese esses modelos também são utilizados nas especificações do modelo de riscos múltiplos com dependência do Capítulo 3 e do modelo de misturas generalizado do Capítulo 4. A Tabela 2.2 mostra as funções de sobrevivência, taxa de falha, densidade e também os parâmetros e momentos desses modelos.

A distribuição exponencial é muito utilizada por sua simplicidade matemática e também por ter propriedades importantes. Sua parametrização é dada pelo parâmetro de escala $\mu > 0$, que equivale à sua esperança, ou por $\lambda = 1/\mu$, que equivale à sua taxa de falha. Uma de suas propriedades é a falta de memória, em que

$$P[T \geq t + z | T \geq t] \quad = \quad P[T \geq z]. \tag{2.25}$$

Essa propriedade faz com que o modelo seja mais tratável, porém limita sua aplicabilidade. Nessa situação, a vida residual média é constante, isto é, $E[X - t|X \geq t] = E[X] = \mu$. Também, a função

de taxa de falha é constante, pois a probabilidade condicional de falha em qualquer tempo t, dado que o evento não ocorreu até t, não depende de t. Essa propriedade de função de taxa de falha é muito restritiva para problemas de análise de sobrevivência e confiabilidade.

A distribuição Weibull tem os parâmetros de escala $\mu > 0$ e de forma $\beta > 0$. Se $\beta < 1$, então a função de taxa de falha do modelo é decrescente; se $\beta > 1$, então a função de taxa de falha é crescente; e se $\beta = 1$, então a distribuição equivale à distribuição exponencial, de taxa de falha constante. Em algumas situações é útil trabalhar com o logaritmo dos tempos de vida. Nesse caso, fazendo $Y = \log X$, em que X tem distribuição Weibull, Y tem distribuição

$$f(y) = \beta \exp\{\beta[y - \log \mu] - \exp\{\beta[y - \log \mu]\}\}, y \in \mathcal{R}, \tag{2.26}$$

que é uma distribuição valor extremo com parâmetros $\mu_0 = \log \mu$ e $\sigma = 1/\beta$. Fazendo $Y = \mu_0 + \sigma W$, segue que W tem distribuição valor extremo padrão com função densidade

$$f(w) = exp\{w - exp^w\}, \tag{2.27}$$

e temos o modelo escrito em forma de modelo linear.

Uma variável aleatória é dita ter distribuição log-normal se seu logaritmo segue uma distribuição normal. Seus parâmetros são o parâmetro de locação $\mu \in \mathcal{R}$ e de forma $\sigma > 0$. A função de risco log-normal $h(x)$ tem forma unimodal, iniciando em zero quando x vale zero, atingindo um valor máximo e retornando a zero conforme x tende a infinito. Esse modelo é criticado como distribuição de tempo de vida por ter taxa de falha decrescente para x grande, o que não é plausível em muitas situações práticas. Nesses casos, o modelo pode ser útil quando os valores grandes de x não são de interesse. Fazendo $Y = \log X = \mu + \sigma W$, temos o modelo na forma de modelos lineares para W com distribuição normal padrão.

Uma variável aleatória X é dita seguir uma distribuição log-logística se seu logaritmo $Y = \log X$ segue uma distribuição logística. Essa distribuição é semelhante à distribuição normal, mas sua função de sobrevivência é mais tratável. Seus parâmetros são o parâmetro de escala $a > 0$ e de forma $b > 0$. Para trabalhar com o logaritmo dos tempos observados, fazemos $Y = \log X$, que tem distribuição logística, com função densidade

$$f(y) = \frac{1}{\sigma} \exp\{\frac{y - \mu}{\sigma}\} \left(1 + \exp\{\frac{y - \mu}{\sigma}\}\right)^{-2}, y \in \mathcal{R} \tag{2.28}$$

com parâmetros de locação e escala dados por $\mu \in \mathcal{R}$ e $\sigma > 0$ e relacionados aos parâmetros a e b por $b = 1/\sigma$ e $a = e^{\mu}$. Essa relação é igual à dada na relação dos parâmetros da distribuição Weibull e valor extremo.

A distribuição gama tem propriedades similares à distribuição Weibull, mas não é igualmente tratável matematicamente. Seus parâmetros são o parâmetro de escala $\mu > 0$ e de forma $k > 0$. A distribuição gama inclui a exponencial com $k = 1$, aproxima a distribuição normal se $k \to \infty$ e é uma distribuição qui-quadrado com ν graus de liberdade se $\nu = 2k$ e $\mu = 2$. Para $k > 1$ a função de taxa de falha é monótona crescente valendo zero em $x = 0$ e tendendo a $1/\mu$ se x tende a infinito. A distribuição gama generalizada inclui um parâmetro a mais no modelo, α, permitindo uma flexibilidade adicional para selecionar uma função de taxa de falha. Sua função densidade é dada por

$$f(x) = \frac{\alpha x^{\alpha k - 1} exp(-x^{\alpha}/\mu)}{\mu^k \Gamma(k)} \tag{2.29}$$

e sobrevivência

$$S(x) = 1 - \frac{\gamma(k, x^\alpha/\mu)}{\Gamma(k)}, \tag{2.30}$$

em que a função $\gamma(.,.)$ está definida na Tabela 2.2. Por incluir as distribuições exponencial, quando $\alpha = k = 1$, lognormal, quando $k \to \infty$, Weibull, quando $k = 1$, e gama quando $\alpha = 1$, ela é uma distribuição muito útil para análise de ajuste de modelos.

2.2.5 Análise de Regressão

Os modelos paramétricos discutidos na seção anterior são úteis para a análise de populações homogêneas. Muitas vezes, porém, os dados de tempos de vida são relativos a populações heterogêneas em que informações adicionais sobre os indivíduos são disponíveis em covariáveis. Também, frequentemente o interesse é conhecer qual é a relação que existe entre as distribuições dos tempos de falha com uma ou mais dessas covariáveis. A seguir discutimos brevemente três métodos básicos de considerar covariáveis nas análises de tempos de vida, sendo uma alternativa paramétrica, de modelos de vida acelerados, e duas alternativas semiparamétricas de modelos de riscos multiplicativos, de Cox, e aditivos, de Aalen.

Considere o tempo de falha $X > 0$ e um vetor $Z' = (Z_1, \cdots, Z_p)$ de variáveis explanatórias associadas com o tempo de falha X. Essas variáveis podem ser quantitativas, qualitativas ou dependentes do tempo. Neste caso, $Z'(t) = (Z_1(t), \cdots, Z_p(t))$.

Duas abordagens de modelagem do efeito de covariáveis na sobrevivência tem se tornado populares na literatura de estatística. A primeira é análoga à abordagem clássica de regressão linear. Nessa abordagem é modelado o logaritmo do tempo de falha $Y = \log X$, que é a mesma transformação que a transformação tradicional utilizada para converter variáveis positivas em variáveis que assumem valores em toda a reta real. O modelo suposto para Y é dado por

$$Y = \mu + \gamma' Z + \sigma W, \tag{2.31}$$

em que $\gamma = (\gamma_1, \cdots, \gamma_p)'$ é um vetor de coeficientes de regressão e W é o erro. Distribuições comuns para o erro incluem a distribuição normal, para que X segue uma distribuição log-normal; valor-extremo, para que X segue uma distribuição Weibull; e logística, para que X segue uma distribuição log-logística. Esse modelo é chamado de modelo de vida acelerado. Para compreender essa nomenclatura, considere $S_0(x)$ a função de sobrevivência de $X = e^Y$ quando Z é zero, o que quer dizer que $S_0(x)$ é a função de sobrevivência de $\exp(Y) = \exp(\mu + \sigma W)$. Mas

$$\begin{aligned} P[X > x|Z] &= P[Y > \log x|Z] \\ &= P[\mu + \gamma' Z + \sigma W > \log x|Z] \\ &= P[\mu + \sigma W > \log x - \gamma' Z|Z] \\ &= P[e^{\mu+\sigma W} > x e^{-\gamma' Z}|Z] \\ &= S_0[x e^{-\gamma' Z}], \end{aligned} \tag{2.32}$$

em que o efeito das variáveis explanatórias é modificar o tempo da escala pelo fator $e^{-\gamma' Z}$. Conforme o sinal de $-\gamma' Z$, o tempo é acelerado ou desacelerado pelo fator. Apesar de o método ser uma extensão direta da análise de regressão, o seu uso para dados de sobrevivência é limitado pelas distribuições que podem ser utilizadas.

A outra abordagem principal é baseada na informação da taxa de falha, a que velocidade indivíduos de certa idade estão passando pelo evento de interesse. Essa abordagem modela a taxa de falha condicional como função das covariáveis. Principalmente, duas classes de modelos são usadas para relacionar efeitos de covariável à sobrevivência, a família de modelos de riscos multiplicativos e a família

de modelos de riscos aditivos.

Para a família de modelos de riscos multiplicativos a taxa de falha condicional de um indivíduo com vetor de covariáveis $\mathbf{z}$ é um produto de uma função de taxa de falha base $h_0(t)$ e uma função não negativa das covariáveis $c(\beta' z)$,

$$h(t|z) = h_0(t)c(\beta' z). \tag{2.33}$$

A função $h_0(t)$ pode ser especificada parametricamente ou permitido ser uma função não negativa. Qualquer função não negativa pode ser usada para $c()$. A maioria das aplicações utiliza o modelo de Cox (1972) com $c(\beta' z) = \exp(\beta' z)$ por sua simplicidade e também por ser uma função positiva para qualquer valor de $\beta' z$. Nessa família, quando todas as covariáveis são fixas no tempo zero, as taxas de falha de dois indivíduos com covariáveis z diferentes, são proporcionais. Por exemplo, para dois indivíduos com covariáveis z_1 e z_2 temos

$$\frac{h(t|z_1)}{h(t|z_2)} = \frac{h_0(t)c(\beta' z_1)}{h_0(t)c(\beta' z_2)} = \frac{c(\beta' z_1)}{c(\beta' z_2)}. \tag{2.34}$$

Por (2.33) temos que a função de sobrevivência condicional de um indivíduo de covariável z pode ser expressa em função de uma função de sobrevivência base $S_0(t)$ como

$$S(t|z) = S_0(t)^{c(\beta' z)}. \tag{2.35}$$

Na família de modelos de riscos aditivos a função taxa de falha condicional é modelada por

$$h(t|z) = h_0(t) + \sum_{j=1}^{p} z_j(t)\beta_j(t). \tag{2.36}$$

Os coeficientes de regressão para esses modelos são funções do tempo de forma que o efeito de uma determinada covariável na sobrevivência é permitido variar com o tempo. As p funções de regressão podem ser positivas ou negativas, mas seus valores são restritos pois (2.36) deve ser positiva. A estimação dos modelos aditivos é feita por mínimos quadrados ponderados.

2.2.6 Riscos Competitivos

Nas seções anteriores consideramos o tempo até a ocorrência do evento de interesse sem levar em conta as diferentes causas possíveis para a ocorrência desse evento, chamadas de riscos competitivos. A situação em que há riscos competitivos é modelada considerando que os indivíduos estão sujeitos à falha por meio de um número k ($k \geq 2$) de causas de falha. Nessa situação somente o tempo mínimo é observado, e a sua ocorrência impede a observação dos tempos devidos às demais causas. Esse tempo de falha é denominado tempo de falha geral. Suas funções de sobrevivência e taxa de falha são denominadas função de sobrevivência geral e função taxa de falha geral.

Uma das formas de construir o modelo de riscos competitivos é por meio de tempos latentes de falha. Kalbfleisch e Prentice (1980) discutem outras maneiras de formalizar o modelo. Consideramos X_j, $j = 1, ..., k$, o tempo potencial não observável de ocorrência do j-ésimo risco competitivo. De cada indivíduo em observação é conhecido o tempo de falha $T = \min\{X_j, j = 1, ..., k\}$ e qual foi a causa da falha por uma variável indicadora $\delta = l$ se $T = X_l$.

A função básica na análise de riscos competitivos é a taxa de falha de causa específica do l-ésimo risco, dada por

$$h_l(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t, \delta = l | T \geq t]}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t, \delta = l | X_j \geq t, j = 1, ..., k]}{\Delta t}. \quad (2.37)$$

Essa função mede a taxa a que sujeitos que não falharam sofrem a l-ésima causa de falha. A taxa de falha geral do tempo T é dada pela soma das k taxas de falha de causa específicas

$$h_T(t) = \sum_{j=1}^{k} h_j(t). \quad (2.38)$$

A taxa de falha de causa específica pode ser derivada a partir da função de sobrevivência conjunta dos k riscos competitivos. Para $S(t_1, \cdots, t_k) = P[X_1 > x_1, \cdots, X_k > x_k]$ a taxa de falha de causa específica é dada por

$$h_l(t) = \frac{-\partial S(t_1, \cdots, t_k)/\partial t_l|_{t_1 = \cdots = t_k = t}}{S(t, \cdots, t)}. \quad (2.39)$$

Para riscos competitivos independentes a taxa de falha de causa específica e a taxa de falha marginal do l-ésimo risco são iguais. Isso não é necessariamente verdade para riscos dependentes. Para ilustrar o caso independente, suponha que os k tempos de falha potenciais são independentes com funções de sobrevivência $S_j(t)$, $j = 1, ..., k$. Segue que sua distribuição conjunta é dada por $S(t_1, \cdots, t_k) = \prod_{j=1}^{k} S_j(t_j)$. Por (2.39), segue que

$$\begin{aligned} h_l(t) &= \frac{-\partial \prod_{j=1}^{k} S_j(t_j)/\partial t_l|_{t_1 = \cdots = t_k = t}}{\prod_{j=1}^{k} S_j(t, \cdots, t)} \\ &= \frac{-\partial S_l(t_l)/\partial t_l|_{t_l = t}}{S_l(t)}. \end{aligned} \quad (2.40)$$

Na modelagem de riscos competitivos frequentemente é necessário fazer suposições sobre a forma de dependência entre os tempos de falha potenciais. Dado que somente é possível observar o tempo e a causa da falha e não todos os tempos de falha potenciais, essas suposições não são testáveis somente pelos dados de riscos competitivos. Esse é o dilema de identificabilidade em riscos competitivos. Para ilustrar essa situação considere $k = 2$ riscos competitivos com função de sobrevivência conjunta dada por $S(t_1, t_2) = [1 + \theta(\lambda_1 t_1 + \lambda_2 t_2)]^{-1/\theta}$, $\theta > 0$ e $\lambda_1, \lambda_2 > 0$. Nesse modelo os dois tempos de falha são correlacionados e apresentam τ de Kendall igual a $\theta/(\theta + 2)$. Por (2.39) temos

$$\begin{aligned} h_l(t) &= \frac{-\partial[1 + \theta(\lambda_1 t_1 + \lambda_2 t_2)]^{-1/\theta}/\partial t_l|_{t_1 = t_2 = t}}{[1 + \theta t(\lambda_1 + \lambda_2)]^{-1/\theta}} \\ &= \frac{\lambda_l}{1 + \theta t(\lambda_1 + \lambda_2)}, l = 1, 2. \end{aligned} \quad (2.41)$$

Nesse caso as funções de sobrevivência marginais, dos tempos latentes X_1 e X_2, são dadas por $S(t, 0) = [1 + \theta t \lambda_1]^{-1/\theta}$ e $S(0, t) = [1 + \theta t \lambda_2]^{-1/\theta}$ o que leva às taxas de falha marginais $\lambda_l/(1 + \theta \lambda_l t)$ que são diferentes das taxas de falha de causa específica. Suponha também dois riscos competitivos independentes com taxas de falha dadas por $\lambda_1/[1 + \theta t(\lambda_1 + \lambda_2)]$ e $\lambda_2/[1 + \theta t(\lambda_1 + \lambda_2)]$. Dada a independência, as taxas de falha de causa específica desses riscos são idênticas às dos riscos dependentes. Isso mostra que os dados observáveis (T, δ) não permitem distinguir um par de riscos competitivos dependentes e independentes.

Em problemas de riscos competitivos muitas vezes o interesse não está na taxa de falha, mas em alguma probabilidade que sumarize o conhecimento disponível sobre a ocorrência de um risco competitivo em particular. É possível calcular três probabilidades diferentes, que são as probabilidades

bruta *(crude)*, líquida *(net)* e bruta parcial *(partial crude)*. A probabilidade bruta é a probabilidade de falha devido a uma causa enquanto as demais causas possíveis estão atuando no indivíduo. A probabilidade líquida é a probabilidade de falha na situação hipotética que um determinado risco é o único atuante na população. Na formulação de tempos latentes a probabilidade líquida equivale à probabilidade marginal da respectiva causa de falha. Probabilidade bruta parcial é a probabilidade de falha numa situação hipotética em que alguns riscos são eliminados.

As probabilidades brutas são frequentemente expressas pela função de incidência acumulada, ou subdistribuição de causa específica, definida por $F_l(t) = P[T \leq t, \delta = l]$. $F_l(t)$ não é uma função de distribuição dado que $F_l(\infty) = P(\delta = l)$. Por suas propriedades de não decrescimento, $F_l(0) = 0$ e $F(\infty) < 1$ ela é uma subdistribuição. $F_l(t)$ pode ser obtida diretamente da função densidade conjunta dos tempos latentes de falha, ou das funções de taxa de falha de causa específica por

$$F_l(t) \quad = \quad P[T \leq t, \delta = l] = \int_0^t h_l(u) \exp(-H_T(u))du, \tag{2.42}$$

em que $H_T(t) = \sum_{j=1}^k \int_0^t h_j(u)du$ é a função taxa de falha acumulada de T. Dessa forma, segundo (2.42), $F_l(t)$ é estimada diretamente dos dados observados, que são os tempos de falha e a taxa de falha de causa específica, sem a necessidade de suposições sobre a função de distribuição conjunta dos tempos latentes.

A função de sobrevivência líquida, ou marginal, $S_l(t)$ é encontrada tomando-se zero para as coordenadas $j = 1, ..., k$, $j \neq l$ na função de sobrevivência conjunta $S(t_1, \cdots, t_k)$. Quando os riscos competitivos são independentes, a função de sobrevivência líquida é relacionada às probabilidades brutas por

$$S_l(t) \quad = \quad \exp\left[\int_0^t \frac{dF_l(u)}{S_T(u)}du\right]. \tag{2.43}$$

Quando os riscos são dependentes, as probabilidades líquidas podem ser cotadas pelas probabilidades brutas. Peterson (1976) mostra que

$$S_T(t) \leq S_l(t) \leq 1 - F_l(t). \tag{2.44}$$

A cota inferior (superior) corresponde à dependência positiva (negativa) perfeita entre os riscos. Em Klein e Moeschberger (1988) e Zheng e Klein (1994) é mostrado que a largura dada por essas cotas pode ser diminuída assumindo uma família de estrutura de dependência para a distribuição conjunta dos riscos competitivos.

Para as probabilidades brutas parciais é definido um conjunto J de causas pelas quais os indivíduos podem falhar e J^c o conjunto de causas que são eliminadas da modelagem. Para $T^J = \min(X_j, j \in J)$ é definida a função de subdistribuição bruta parcial $F_l^J(t) = P[T^J \leq t, \delta = l]$, $l \in J$. A l-ésima probabilidade bruta parcial é a chance de falha devido à causa l para uma unidade de observação que somente pode sofrer a falha devido às causas em J. A taxa de falha bruta parcial é relacionada à função de sobrevivência conjunta por

$$\lambda_l(t) = \frac{-\partial S(t_1, \cdots, t_k)/\partial t_l|_{t_j=t, t_j \in J, t_j=0, t_j \in J^c}}{S(t_1, \cdots, t_k)|_{t_j=t, t_j \in J, t_j=0, t_j \in J^c}}, \tag{2.45}$$

e a função de subdistribuição bruta parcial por

$$F_l^J(t) \quad = \quad P[T^J \leq t, \delta = l] = \int_0^t \lambda_l^J(x) \exp\left(-\sum_{j \in J} \int_0^t \lambda_j^J(u)du.\right) dx. \tag{2.46}$$

Quando os riscos competitivos são independentes, as taxas de falha brutas parciais podem ser expressas em termos das probabilidades brutas por

$$\lambda_l^J(t) = \frac{dF_l(t)/dt}{S_T(t)}. \tag{2.47}$$

2.3 Misturas Finitas

Os modelos de misturas de distribuições exibem grande flexibilidade para modelagem de dados e por isso são amplamente utilizados na literatura em diversos campos de aplicação como astronomia, biologia, genética, psiquiatria, engenharia, marketing etc. Nessas aplicações os modelos de misturas dão sustentação a diversas técnicas estatísticas, entre elas os modelos de estruturas latentes, investigados nessa tese. Titterington *et al.* (1985) e Mclachlan e Peel (2000) são dedicados a modelos de misturas finitas.

O modelo de misturas finitas univariado de g componentes é definido como

$$f(t;\Upsilon) = \sum_{j=1}^{g} p_j f_j(t;\Upsilon_j), \tag{2.48}$$

em que $p_j \in (0,1)$ são constantes satisfazendo $\sum_{j=1}^{g} p_j = 1$, $f_j(t;\Upsilon_j)$, $j = 1,..,g$ são funções de densidade de probabilidade com parâmetros desconhecidos Υ_j, sendo que $\Upsilon = (\Upsilon_1, \cdots, \Upsilon_g)$. $f_j(t)$, $j = 1,..,g$, são chamadas de densidades componentes da mistura. Uma das interpretações dos modelos de misturas é a presença de g grupos heterogêneos nos dados, seguindo as proporções $p_1, ..., p_g$.

A abordagem de misturas de distribuições tem a flexibilidade de abordagens não paramétricas ao mesmo tempo que as vantagens da abordagem paramétrica como, por exemplo, manter a dimensão do espaço paramétrico em nível razoável. Isso torna os modelos de misturas uma alternativa importante de estimação de densidades. Para ilustrar essa situação, Priebe (1994) mostrou que com $n = 10.000$ observações uma distribuição log-normal pode ser bem aproximada por uma mistura de 30 distribuições normais, enquanto que um estimador de densidade de núcleo utiliza uma mistura de $n = 10.000$ distribuições normais. Também, Mclachlan e Peel (2000) ilustram a construção de modelos com várias formas assimétricas e contendo multimodalidade a partir de misturas de distribuições normais, controlando pelos pesos das componentes e pelas distâncias entre suas modas.

Um exemplo de modelagem que utiliza misturas de distribuições é o modelo de fração de cura, ou taxa de cura. Ele foi inicialmente proposto por Boag (1949) e depois desenvolvido por Berkson e Gage (1952). Nesse modelo de sobrevivência uma proporção da população é suposta não passar pela experiência de falha e o restante segue um modelo de sobrevivência usual. Seja $S_{pop}(t|Z_i, X_i)$ a função de sobrevivência da população, que é imprópria devido à parcela de indivíduos curados da população, isto é, $S_{pop}(\infty|Z_i, X_i) > 0$. X_i e Z_i são vetores de covariáveis. Seja também $S(t|X_i)$ a função de sobrevivência dos indivíduos não curados, que é própria, isto é, $S(\infty|X_i) = 0$. O modelo de misturas de cura de Berkson e Gage (1952) é a mistura de uma fração $1 - \gamma(Z_i)$ da população, que é curada, e o restante de proporção $\gamma(Z_i)$ da população, que é não curada, de forma que

$$S_{pop}(t|Z_i, X_i) = 1 - \gamma(Z_i) + \gamma(Z_i)S(t|X_i). \tag{2.49}$$

À probabilidade $\gamma(Z_i)$ podem ser associadas covariáveis Z, por exemplo, utilizando regressão logística

$$\gamma(Z_i) = \frac{\exp(Z_i'\beta)}{1 + \exp(Z_i'\beta)}, \tag{2.50}$$

em que β é um vetor de coeficientes de regressão.

Tabela 2.1: *Funções de cópula utilizadas nesse trabalho. $D_k(x)$ denota a função Debye $k/x^k \int_0^x \frac{t^k}{e^t-1}dt$, $k = 1, 2$. Para as cópulas de Gumbel, Clayton e Joe-Clayton simetrizada, a medida de dependência ρ de Spearman é dada por $\rho = 12 \int \int_{[0,1]^2} C(u, v)dudv - 3$. C_{JC} denota a cópula de Joe-Clayton, dada por $C_{JC}(u, v) = 1 - \left[\{[1 - (1 - u)^\kappa]^{-\gamma} + [1 - (1 - v)^\kappa]^{-\gamma} - 1\}^{-1/\gamma}\right]^{1/\kappa}$.*

Cópula	$C(u, v)$	Domínio de θ	τ de Kendall	ρ de Spearman
Produto	uv	$-$	0	0
Clayton	$(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$	$\theta \in (0, \infty)$	$\frac{\theta}{\theta+2}$	$*$
Gumbel	$\exp(-(\tilde{u}^\theta + \tilde{v}^\theta)^{1/\theta})$, $\tilde{x} = -\log(x)$	$\theta \in (1, \infty)$	$\frac{\theta-1}{\theta}$	$*$
Frank	$-\log\left(1 - \frac{(1-e^{-\theta u})(1-e^{-\theta v})}{1-e^{-\theta}}\right)/\theta$	$\theta \in (-\infty, \infty)$	$1 - \frac{4}{\theta}[1 - D_1(\theta)]$	$1 - \frac{12}{\theta}[D_1(\theta) - D_2(\theta)]$
Joe-Clayton Simetrizada	$\frac{C_{JC}(u,v\|\tau_U,\tau_L)}{2} + \frac{C_{JC}(1-u,1-v\|\tau_L,\tau_U)}{2} + u + v - 1$	$\theta = (\kappa, \gamma) \in (0, \infty)^2$, $\kappa = \frac{1}{\log_2(2-\tau_U)}$, $\gamma = \frac{-1}{\log_2(\tau_L)}$	$*$	$*$

Tabela 2.2: *Funções de sobrevivência, taxa de falha e densidade de modelos paramétricos de análise de sobrevivência. As funções gama completa e gama incompleta são dadas, respectivamente, por* $\Gamma(x) = \int_0^\infty t^{x-1}exp(-t)dt$ *e* $\gamma(k, t/\mu) = \int_0^\infty (t/\mu)^{k-1}exp(-t/\mu)/\mu dt$. z_p *é o percentil de 100p% da distribuição normal padrão.*

Distribuição Parâmetros	S(t)	h(t)	f(t)	Quantil-p	E[X]	V[X]
Exponencial $\mu > 0$	$e^{-t/\mu}$	$1/\mu$	$e^{-t/\mu}/\mu$	$-\mu \log(1-p)$	μ	μ^2
Weibull $\mu, \beta > 0$	$e^{-(t/\mu)^\beta}$	$\frac{\beta}{\mu}(t/\mu)^{\beta-1}$	$\frac{\beta}{\mu}(t/\mu)^{\beta-1}e^{-(t/\mu)^\beta}$	$-\mu\,[ln(1-p)]^\beta$	$\mu\Gamma(1+\frac{1}{\beta})$	$\mu^2\Gamma(1+\frac{2}{\beta}) - E[X]^2$
Log-normal $\sigma > 0, \mu \in \mathcal{R}$	$1 - \Phi\left[\frac{lnt-\mu}{\sigma}\right]$	$\frac{f(t)}{S(t)}$	$\frac{exp\left[-\frac{1}{2}(\frac{lnt-\mu}{\sigma})^2\right]}{x(2\pi)^{1/2}\sigma}$	$exp(\mu + \sigma^2 z_p)$	$exp(\mu + \frac{1}{2}\sigma^2)$	$exp(2\mu + 2\sigma^2) - E[X]^2$
Log-logísitca $a, b > 0$	$\left[1 + (x/a)^b\right]^{-1}$	$\frac{b/a(x/a)^{b-1}}{1+(x/a)^b}$	$\frac{b/a(x/a)^{b-1}}{\left[1+(x/a)^b\right]^2}$	$a\left[\frac{p}{1-p}\right]^b$	$\frac{a\pi b}{sen(b)}$, b>1	$\frac{a^2 2\pi}{bsen(2\pi/b)} - E[X]^2$, b>2
Gama $\mu, k > 0$	$1 - \frac{\gamma(k,x/\mu)}{\Gamma(k)}$	$\frac{f(t)}{S(t)}$	$\frac{t^{k-1}exp(-t/\mu)}{\Gamma(k)\mu^k}$		$k\mu$	$k\mu^2$

Capítulo 3

Modelo de Riscos Múltiplos com Dependência

Neste capítulo apresentaremos um resumo dos resultados de dois artigos relacionados ao modelo de riscos múltiplos com dependência. O primeiro, apresentado no Apêndice A e intitulado 'Polyhazard Models With Dependent Causes', foi aceito para publicação no Brazilian Journal of Probability and Statistics. O segundo, 'Fitting distributions with the polyhazard model with dependence', é apresentado no Apêndice B e foi aceito para publicação na revista Communications in Statistics - Theory and Methods.

O modelo de riscos múltiplos com dependência generaliza, por meio de cópulas, os modelos de riscos múltiplos independentes de forma a permitir que haja dependência entre as causas latentes de falha. A família de modelos dada por essa generalização é capaz de representar várias formas de dependência e também permite a especificação de qualquer distribuição marginal para os tempos latentes, resultando em modelos capazes de gerar funções de taxa de falha muito mais flexíveis que os modelos de riscos múltiplos independentes, incluindo formas de banheira, multimodalidade e efeitos locais dados pelos riscos competitivos. Nesse capítulo utilizamos os termos riscos dependentes ou riscos independentes, para fazer diferença entre os modelos de riscos múltiplos dependentes e independentes.

3.1 O Modelo de Riscos Múltiplos Independentes

A principal vantagem dos modelos de riscos múltiplos em comparação com os modelos de risco simples é a flexibilidade de representar funções de taxa de falha com formas não usuais. Há vários exemplos de modelos de riscos múltiplos independentes na literatura. Kalbfleisch e Prentice (1980) propuseram o modelo poli-log-logístico para riscos competitivos log-logísticos. Berger e Sun (1993) propuseram o modelo poli-Weibull para riscos competitivos Weibull e sua estimação por análise Bayesiana usando o amostrador de Gibbs. Louzada-Neto (1999) propôs um modelo de riscos múltiplos generalizado que inclui os modelos poli-Weibull, poli-log-logístico e poli-gama-generalizada. Também propôs a inclusão de covariáveis em uma forma híbrida de modelo de vida acelerado e de riscos proporcionais. Kuo e Yang (2000) e Basu *et al.* (1999) usaram o modelo poli-Weibull para modelar sistemas encobertos *(masked systems)*, nos quais a causa de falha pode ser desconhecida ou parcialmente conhecida. Mazucheli *et al.* (2001) propuseram um procedimento de inferência Bayesiana para os modelos poli-Weibull e poli-log-logístico com covariáveis. Louzada-Neto *et al.* (2004) analisaram a identificabilidade na estimação do modelo poli-Weibull pelas matrizes de informação e correlação dos estimadores dos parâmetros e também por um teste da hipótese de não identificabilidade baseado em *bootstrap*.

O modelo de riscos múltiplos independentes é baseado na situação de riscos competitivos em que a causa de falha não é conhecida. Considere n unidades em observação e $k \geq 2$ causas latentes

de falha exercendo influência sobre cada uma dessas unidades. Suponha que o tempo de vida relacionado à j-ésima causa de falha da i-ésima unidade em observação, X_{ij}, tenha função densidade $f_j(\cdot;\Gamma_j)$, considerada conhecida exceto pelo conjunto de parâmetros desconhecidos Γ_j. Denote as funções de sobrevivência e de taxa de falha de X_{ij} por $S_j(\cdot;\Gamma_j)$ e $\lambda_j(\cdot;\Gamma_j)$, respectivamente. Somente $X_i = \min\{X_{ij}, j = 1,\cdots,k\}$ é observado de cada unidade em observação. Portanto, considerando a independência entre os riscos, isto é, entre os tempos de falha X_{ij}, $j = 1,\cdots,k$, a função de sobrevivência geral de X_i, denotada por $S(t;\Upsilon)$, em que $\Upsilon = (\Gamma_1,\cdots,\Gamma_k)$, é dada para qualquer $i = 1,\cdots,n$ pelo produto das funções de sobrevivência marginais, isto é,

$$\begin{aligned} S(t;\Upsilon) &= P_\Upsilon[X_i > t] \\ &= P_\Upsilon[X_{i1} > t,\cdots,X_{ik} > t] \\ &= \prod_{j=1}^{k} S_j(t;\Gamma_j), \end{aligned} \tag{3.1}$$

do que segue que a função taxa de falha de X_i, $\lambda(t;\Upsilon)$, é dada pela soma das taxas de falha marginais, pois

$$\begin{aligned} \lambda(t;\Upsilon) &= -\frac{d}{dt}\log S(t;\Upsilon) \\ &= -\frac{d}{dt}\sum_{j=1}^{k}\log S_j(t;\Gamma_J) \\ &= \sum_{j=1}^{k}\lambda_j(t;\Gamma_j). \end{aligned} \tag{3.2}$$

Note que nessa modelagem, de forma diferente da situação clássica de riscos competitivos, a informação da causa de falha de cada observação não é disponível, o que torna essa situação diferente da situação de riscos competitivos clássica. Essa maior escassez de informação faz com que não sejam observáveis as taxas de falha de causa específica, discutidas no Capítulo 2, mas apenas a taxa de falha geral. As questões de identificabilidade do modelo são discutidas mais adiante neste capítulo.

3.2 O Modelo de Riscos Múltiplos com Dependência

No artigo (Apêndice A) a dependência entre os tempos latentes de falha é modelada usando funções de cópulas. Conforme discutido no Capítulo 2 uma função de cópula m-dimensional pode ser definida como uma função de distribuição acumulada cujas distribuições marginais são uniformes em $[0,1]$ e cujo domínio é o hipercubo $[0,1]^m$. Funções de cópula têm sido bastante estudadas na literatura de modelagem multivariada, principalmente quando o uso da distribuição normal é questionável. Uma vantagem importante da abordagem de cópulas é a possibilidade de modelar a dependência e o comportamento marginal das variáveis relacionadas separadamente, o que faz da cópula uma alternativa bastante prática no caso de modelagem multivariada. Algumas referências para cópulas incluem os livros Nelsen (2006), Joe (1997) e Cherubini *et al.* (2004) e o artigo de Trivedi e Zimmer (2005). Em análise de sobrevivência problemas multivariados são tratados com abordagem de cópulas, por exemplo, em Clayton (1978), Oakes (1982), Georges *et al.* (2001) e Romeo *et al.* (2006).

Devido à modelagem de tempos de vida dos modelos de riscos múltiplos, é natural que a função de cópulas modele o comportamento conjunto de sobrevivência dos tempos latentes ao invés do comportamento desses tempos segundo a função de distribuição. Para isso, o problema de considerar dependência entre os riscos é formulado em termos da função de sobrevivência conjunta e os resultados

de cópulas mais popularmente utilizados para funções de distribuição são adaptados para funções de sobrevivência.

Deste ponto em diante consideramos notação para $k = 2$, que pode ser diretamente generalizada para $k > 2$. Denotando por $H(\cdot,\cdot;\Upsilon)$ a função de distribuição conjunta e por $\bar{H}(\cdot,\cdot;\Upsilon)$ a função de sobrevivência conjunta das variáveis latentes X_{i1} e X_{i2}, com $\Upsilon = (\Gamma_1, \Gamma_2)$, podemos escrever a função de sobrevivência de X_i como

$$\begin{aligned} S(t;\Upsilon) &= P_\Upsilon[X_{i1} > t, X_{i2} > t] \\ &= \bar{H}(t,t;\Upsilon) \end{aligned} \tag{3.3}$$

e, portanto, o objetivo é considerar a modelagem de cópulas da função de sobrevivência conjunta $\bar{H}$ dos tempos latentes aplicada ao problema de riscos competitivos com causa de falha desconhecida, dada em (3.3).

Para $F_1(\cdot;\Gamma_1)$ e $F_2(\cdot;\Gamma_2)$, as funções de distribuição de X_{i1} e X_{i2}, respectivamente, segue pelo teorema de Sklar que existe uma função de cópula C^* tal que podemos escrever $H(t_1,t_2;\Upsilon)$=$C^*(F_1(t_1;\Gamma_1), F_2(t_2;\Gamma_2))$ e que, C^* é única se as distribuições marginais F_1 e F_2 são contínuas. C^* é então chamada de função de cópula, pois ela acopla as distribuições marginais F_1 e F_2 à sua distribuição conjunta H.

É possível representar a função de sobrevivência conjunta diretamente por $\bar{H}(t_1,t_2;\Upsilon)$=$P[X_1 > t_1, X_2 > t_2;\Upsilon]$=$\tilde{C}(S_1(t_1;\Gamma_1), S_2(t_2;\Gamma_2))$, em que $\tilde{C}(u,v) = u+v-1+C^*(1-u,1-v)$ também é uma cópula, conforme discutido no Capítulo 2. Por outro lado, para qualquer cópula C, $C(S_1(t_1;\Gamma_1), S_2(t_2;\Gamma_2))$ é uma função de sobrevivência conjunta. Portanto, a função de sobrevivência S também pode ser modelada diretamente pela função de cópula C como em Kaishev *et al.* (2007). Essa também é a abordagem adotada no artigo (Apêndice A), pois geralmente é mais fácil trabalhar analiticamente com essa representação.

Segue que para a função de sobrevivência do modelo de riscos múltiplos com dependência dada por uma função de cópulas C com parâmetro de dependência θ e $\Upsilon = (\theta, \Gamma_1, \Gamma_2)$, podemos escrever

$$\begin{aligned} S(t;\Upsilon) &= \bar{H}(t,t;\Upsilon) \\ &= C_\theta(S_1(t;\Gamma_1), S_2(t;\Gamma_2)), \end{aligned} \tag{3.4}$$

em que S_1 e S_2 são funções de sobrevivência marginais contínuas. A cópula C em (3.4) é chamada cópula de sobrevivência, a que, nesse trabalho, nos referimos simplesmente por função de cópula. Note que, devido à aplicação da cópula às funções de sobrevivência marginais, a dependência na cauda à direita (esquerda) para os tempos latentes de sobrevivência é igual à dependência na cauda à esquerda (direita) da cópula C de (3.4). Da função de sobrevivência (3.4), segue que as funções densidade de probabilidade e de taxa de falha do modelo de riscos múltiplos dependentes são obtidas pela forma usual, isto é,

$$f(t;\Upsilon) = -\frac{d}{dt}S(t;\Upsilon) \quad \text{e} \quad h(t;\Upsilon) = \frac{f(t;\Upsilon)}{S(t;\Upsilon)}. \tag{3.5}$$

O modelo proposto é uma generalização do modelo de riscos múltiplos independentes em que é permitida a dependência concomitantemente à modelagem do comportamento marginal dos riscos competitivos latentes. Para cada combinação de cópula e funções de sobrevivência marginais especificada é construído um novo modelo. Isso possibilita a construção de uma família rica de modelos de riscos competitivos latentes. Por exemplo, no artigo (Apêndice A) são consideradas para as causas de falha latentes as distribuições exponential, log-logistica, log-normal, gama e Weibull e para a dependência as cópulas de Clayton, Gumbel e Frank. Apesar disso, poderiam ser utilizadas quaisquer distribuições marginais e função de cópulas na especificação do modelo. A cópula de Joe-Clayton simetrizada (SJC) também é utilizada nos exemplos do artigo. Essas funções de cópula foram selecionadas por serem

amplamente utilizadas na literatura e exibirem diferentes tipos de dependência. A cópula de Frank, com parâmetro $\theta \in (-\infty, +\infty)$, é uma cópula Arquimediana simétrica com τ de Kendall em $(-1, 1)$ e ρ de Spearman em $(-1, 1)$, e com dependência nas caudas à esquerda e à direita, λ_L e λ_U, iguais a zero. A cópula pode gerar distribuições com dependência forte próximo à região central do seu suporte, mas a dependência nas caudas é sempre baixa. Então, na região das caudas, a função taxa de falha do modelo de riscos múltiplos é sempre aproximadamente igual à soma das funções de taxa de falha marginais. Para a cópula de Clayton, temos o parâmetro $\theta \in (0, +\infty)$, $\tau = \theta/(\theta + 2) \in [0, 1)$, $\rho \in [0, 1)$, $\lambda_U = 2^{-1/\theta} \in (0, 1)$, e $\lambda_L = 0$. Para a cópula de Gumbel, temos o parâmetro $\theta \in [1, +\infty)$, $\tau = (\theta - 1)/\theta \in [0, 1)$, $\rho \in [0, 1)$, $\lambda_U = 0$, e $\lambda_L = 2 - 2^{1/\theta} \in [0, 1)$. Para a cópula SJC, a dependência nas caudas inferior, λ_L, e superior, λ_U, pertencem a $[0, 1)$. Essas características devem ser levadas em consideração para a seleção da função de cópulas (vide Trivedi e Zimmer (2005) e Nelsen (2006) para mais propriedades). Uma motivação para a modelagem com uma faixa ampla para a dependência da cópula é dada em Yashin *et al.* (1986). Eles discutem que os mecanismos das doenças atuantes em indivíduos podem interagir de formas diferentes correspondendo a relações de dependência positiva, em que os mecanismos se potencializam mutuamente, e também, de dependência negativa. Por exemplo, em idade avançada, as condições de redução do metabolismo e ineficiência circulatória podem retardar o crescimento de tumores sólidos.

Para um exemplo de especificação do modelo de riscos múltiplos com dependência, considere a cópula de Frank e tempos de falha latentes Weibull, i.e. $X_{ij} \sim \mathrm{W}(\mu_j; \beta_j)$, $j = 1, 2$, em que μ é parâmetro de escala e β, de forma. Esse modelo é chamado de Frank-Weibull-Weibull, em que o primeiro nome se refere à função de cópula e os dois últimos, às distribuições latentes. De acordo com a notação do modelo proposto, seus parâmetros podem ser denotados por $\Upsilon = (\theta, \Gamma_1, \Gamma_2)$, em que $\Gamma_1 = (\mu_1, \beta_1)$ e $\Gamma_2 = (\mu_2, \beta_2)$. A função de sobrevivência geral de X_i é dada por

$$S(t; \Upsilon) = -\frac{1}{\theta} \log\left(1 - \frac{(1 - e^{-\theta e^{-(t/\mu_1)^{\beta_1}}})(1 - e^{-\theta e^{-(t/\mu_2)^{\beta_2}}})}{(1 - e^{-\theta})}\right), \tag{3.6}$$

e a função densidade de probabilidade de X_i por

$$f(t; \Upsilon) = \frac{(1 - e^{-\theta S_2(t)})e^{-\theta S_1(t)} f_1(t) + (1 - e^{-\theta S_1(t)})e^{-\theta S_2(t)} f_2(t)}{(1 - e^{-\theta}) - (1 - e^{-\theta S_1(t)})(1 - e^{-\theta S_2(t)})}, \tag{3.7}$$

em que f_1 e f_2 são as funções de densidade de X_{i1} e de X_{i2}, respectivamente.

A Figura 3.1, copiada do Apêndice A, ilustra algumas formas possíveis para a distribuição de X_i com especificação Frank-Weibull-Weibull, considerando $X_{i1} \sim W(4,0; 0,9)$ e $X_{i2} \sim W(5,0; 3,0)$, e o parâmetro de dependência variando de forma tal que o τ de Kendall varia de $-0,80$ a $0,80$. A figura mostra as diferentes formas que a função taxa de falha pode assumir, dependendo das formas da distribuição marginal e o tipo de dependência. A Figura 3.2 mostra algumas formas possíveis da função taxa de falha para outras especificações do modelo, em que é possível notar efeitos locais, formas de banheira e multimodal. Os modelos são mostrados de acordo com a dependência entre as variáveis latentes em termos do τ de Kendall, com exceção da cópula SJC em que os valores denotam a dependência de cauda inferior e superior.

Para simplificar a referência às especificações do modelo de riscos múltiplos, desse ponto em diante são utilizadas siglas Indep, Exp, Gam, Llog, Lnor e Wei para a cópula independência, e distribuições exponencial, gama, log-logística, log-normal e Weibull. Por exemplo, a especificação Indep-Lnor-Wei significa o modelo de riscos múltiplos com cópula independência e distribuições marginais log-normal e Weibull.

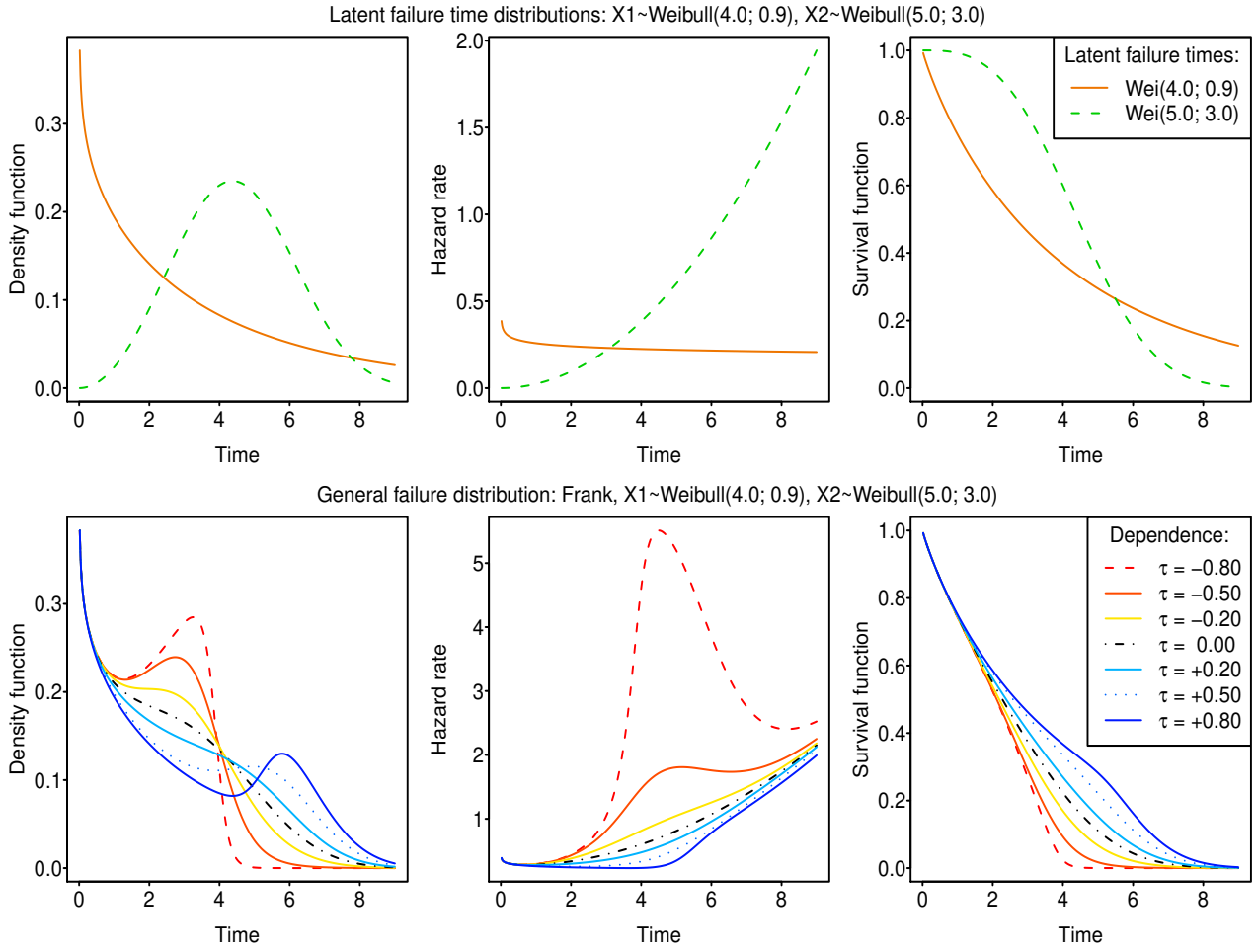


Figura 3.1: *Exemplos de funções densidade, taxa de falha e sobrevivência para o modelo de riscos simples Weibull (superior) e para o modelo de riscos múltiplos com marginais Weibull utilizando dependência segundo uma cópula de Frank (inferior).*

3.3 Identificabilidade do Modelo

Duas discussões são necessárias no assunto de identificabilidade do modelo de riscos múltiplos. A primeira é a identificabilidade dos seus parâmetros devido às formas paramétricas participantes da especificação do modelo. Nesse sentido, alguns modelos são claramente não identificáveis. Considere, por exemplo, o modelo Gumbel-Wei-Wei. A cópula de Gumbel é dada por

$$C(u,v) = \exp[-\{(-\log u)^\theta + (-\log v)^\theta\}^{\frac{1}{\theta}}], \quad u,v \in [0,1].$$

Portanto, por (3.4), considerando distribuições marginais Weibull com parâmetros (λ_1, β_1) e (λ_2, β_2), a função de sobrevivência geral é dada por

$$\begin{aligned} S(t) &= C(S_1(t), S_2(t)) \\ &= \exp[-\{\lambda_1^\theta t^{\theta\beta_1} + \lambda_2^\theta t^{\theta\beta_2}\}^{\frac{1}{\theta}}], \end{aligned}$$

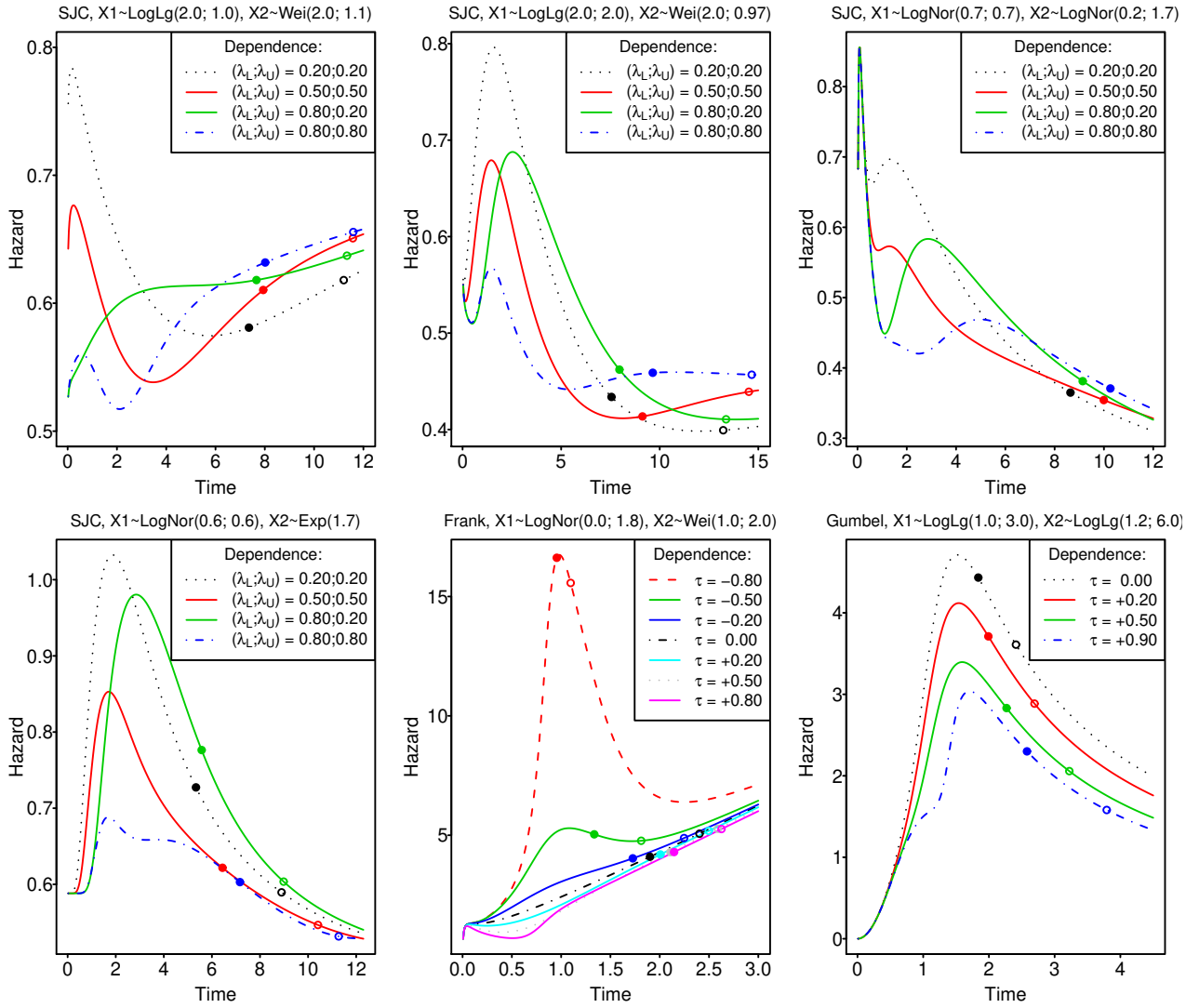


Figura 3.2: *Exemplos de funções taxa de falha para o modelo de riscos múltiplos com dependência. Os dois pontos na figura marcam os quantis de 99% e 99.9% de probabilidade para cada modelo.*

o que mostra que o modelo não é identificável quando $\beta_1 = \beta_2 = \beta$, em que qualquer terna $(\lambda_1', \lambda_2', \theta')$ satisfazendo $(\lambda_1'^{\theta'} + \lambda_2'^{\theta'})^{1/\theta'} = (\lambda_1^{\theta} + \lambda_2^{\theta})^{1/\theta}$ pode gerar o mesmo modelo. O mesmo problema de não identificabilidade ocorre nas subclasses do modelo Gumbel-Wei-Wei: Gumbel-Exp-Exp, Indep-Exp-Exp, e Indep-Wei-Wei. Outro exemplo de não identificabilidade é o modelo Clayton-Llog-Llog na situação em que ambas distribuições marginais têm o mesmo parâmetro de forma e o parâmetro de dependência vale 1. Também temos não identificabilidade quando a distribuição de uma variável latente é dominada estocasticamente pela outra variável no problema de riscos competitivos quando é utilizada uma cópula com dependência positiva perfeita. Analiticamente, o procedimento de verificar a identificabilidade ou não de diversas especificações do modelo de riscos múltiplos é uma tarefa complicada, de forma que a identificabilidade dos demais modelos dados pelas combinações de cópulas de Clayton, Gumbel e Frank com distribuições de tempos latentes exponencial, gama, log-logística, lognormal e Weibull foi realizada por meio de análises numéricas buscando no espaço paramétrico situações em que mais de uma parametrização representando o mesmo modelo. Os resultados desses procedimentos, descritos no

artigo (Apêndice A), detectaram como não identificáveis os casos mencionados acima.

Um outro problema relacionado à identificabilidade no modelo de riscos múltiplos é a identificabilidade dos modelos latentes. Na literatura tradicional de riscos competitivos, quando a causa de falha é conhecida, essa questão é denominada dilema de identificabilidade, conforme ilustrado no Capítulo 2. Essa situação também é abordada em Cox (1972) e Tsiatis (1975). Nesse problema clássico, um modelo de riscos competitivos é identificável se a função de sobrevivência conjunta pode ser calculada ou identificada pelo simples conhecimento da função de sobrevivência geral. Tsiatis (1975) mostrou que, para um modelo com riscos dependentes, é possível encontrar um conjunto de riscos independentes que produz a mesma função de sobrevivência conjunta. Segue que, a menos que restrições sejam impostas ao comportamento dos riscos competitivos, esse tipo de identificabilidade não é possível. Alguns artigos mostram resultados nessa direção. Heckman e Honoré (1989) usam uma função similar a uma cópula baseada em covariáveis para contornar não parametricamente o problema de identificabilidade. Carriere (1994) relaciona as probabilidades brutas marginais às probabilidades líquidas usando funções de cópulas quando existe dependência entre os riscos. Zheng e Klein (1995) mostram que a identificabilidade das distribuições marginais é possível se a função de cópula é completamente determinada.

O modelo de riscos múltiplos pode ser visto como um modelo de riscos competitivos com valores desconhecidos para a causa de falha. Devido à situação de haver menos informação disponível, a identificabilidade do modelo de riscos competitivos equivalente é uma condição necessária mas não suficiente para a identificabilidade do modelo de riscos múltiplos. Porém, mesmo quando temos esse tipo de não identificabilidade nos modelos de riscos múltiplos, essa modelagem ainda é bastante útil devido às boas características do modelo.

3.4 Estimação e Exemplos de Aplicação

A estimação do modelo descrita no artigo (Apêndice A) foi conduzida por máxima verossimilhança, por meio da otimização da função de log-verossimilhança. Em todas as aplicações a estimação foi realizada a partir de cerca de 200 valores iniciais a fim de constatar a existência de problemas de máximos locais e identificabilidade. Exceto pelos problemas de máximos locais observados na estimação das especificações de cópulas, não foram encontrados problemas de convergência nas aplicações aos dados empíricos e simulados.

A análise da matriz Hessiana mostra que para algumas especificações, um número grande de observações é necessário para atngir uma pequena variância do estimador do parâmetro da cópula. Isso é importante quando a diferença entre os modelos de riscos múltiplos dependente e independente ocorre numa região de pouca probabilidade. Isso também é esperado pois é necessário um grande número de observações da falha geral para que seja obtido um número razoável de observações nessa região de probabilidade baixa.

No artigo (Apêndice A) as aplicações foram realizadas para dados simulados, usando dois modelos, Clayton-Llog-Lnor e Frank-Lnor-Wei, e para dados reais sobre duração de desemprego na Alemanha. Para cada conjunto de dados foram ajustadas as especificações resultantes das combinações de cópula de Frank, Clayton, Gumbel e independência e distribuições marginais exponencial, gama, log-logística, log-normal e Weibull e também os modelos de riscos simples dados por essas distribuições para comparação. Os modelos de riscos simples não puderam representar os dados simulados e de duração de desemprego, enquanto que o modelo proposto exibiu um ajuste muito bom aos dados, representando os efeitos resultantes da presença dos riscos competitivos. Apesar de não ter sido possível fazer inferências para os tempos latentes, devido ao problema de identificabilidade resultante da falta de informação da causa de falha, o modelo proposto permite que sejam impostas restrições na dependência (negativa, positiva ou de cauda), e também permite o exame direto da associação entre covariáveis e o comportamento de

tempos latentes.

3.5 Ajustando Distribuições com o Modelo de Riscos Múltiplos Dependentes

Uma versão estendida do modelo de riscos múltiplos com dependência é a que considera marginais com suporte em toda a reta real. Essa proposta de extensão do modelo é analisada no artigo 'Fitting distributions with the polyhazard model with dependence', apresentado no Apêndice B dessa tese. Nesta seção apresentamos um breve resumo do artigo.

O modelo de riscos múltiplos com dependência foi inicialmente proposto para ajustar dados de tempos de vida e nessa situação as distribuições marginais têm suporte nos reais positivos. Ao eliminar essa restrição o método produz uma família rica de distribuições univariadas com assimetria e múltiplas modas. O modelo é capaz de aproximar outras distribuições propostas na literatura e também exibe bom ajuste em comparação com essas metodologias.

A família de modelos tomada como referência nessa análise é a família das distribuições geradas de beta generalizadas (GBG) (Alexander *et al.*, 2011; Cordeiro *et al.*, 2012). As distribuições GBG são definidas pela distribuição beta generalizada do primeiro tipo e por uma distribuição geradora $F(x;\Theta)$ e sua função densidade correspondente $f(x;\Theta)$, para as quais Θ é um vetor de parâmetros (Alexander *et al.* (2012); vide também o Apêndice B). Para cada distribuição geradora, é obtida uma distribuição GBG. Nesse trabalho, foram consideradas as seguintes distribuições da família GBG: Weibull (GBW), normal (GBN), log-normal (GBLN), gama (GBGam), Gumbel (GBGum), t assimétrica com dois graus de liberdade (GBTSk), e Laplace escalada (GBSLa). Também foram consideradas as distribuições beta modificada Weibull (BMW) (Nadarajah *et al.*, 2011), e a beta-gama generalizada (BGGam) de Cordeiro *et al.* (2012). Consideramos o conjunto de distribuições dados pelo método de integração da distribuição beta, isto é, os modelos GBG, BMW e BGGam como modelos gerados de beta (BG).

Ao eliminar a restrição de que as variáveis latentes são positivas, não há mais sentido em considerar distribuições de tempos de vida, e conceitos de sobrevivência também não são mais válidos. Por outro lado, é definido um novo método de gerar uma família rica de distribuições. Utilizando procedimentos de otimização em que modelos de riscos múltiplos com dependência (PHD, *polyhazard with dependent causes*) visam representar casos de modelos BG, é mostrado em particular que o modelo PHD pode aproximar densidades das famílias de distribuições BG. A Figura 3.3, copiada do artigo no Apêndice B, mostra o ajuste de funções de taxa de falha da família PHD a funções de taxa de falha da família BG. Também, a família de densidades PHD exibe bom ajuste a três conjuntos de dados encontrados na literatura. As Figuras 3.4 e 3.5, trazidas do Apêndice B, ilustram esses ajustes. A riqueza de formas de representação da família PHD pode ser aumentada utilizando funções de cópula de dois parâmetros, aumentando o número de variáveis latentes ou usando distribuições da família BG como distribuições latentes.

Nas análises realizadas, os modelos GBTSk e GBSLa da família BG têm quatro parâmetros e os demais modelos utilizados, cinco parâmetros. Os modelos da família PHD utilizados foram os dados pela combinação das cópulas Frank, Gumbel, Clayton, Independência e Joe-Clayton simetrizada e distribuições marginais exponencial, gama, log-logística, log-normal e Weibull. Portanto, os modelos PHD nas aplicações realizadas têm de três a seis parâmetros, contando-se, para a cópula, dois parâmetros no caso da cópula de Joe-Clayton simetrizada, nenhum parâmetro no caso da independência e um parâmetro nos casos das demais cópulas e, para as distribuições marginais, dois parâmetros por distribuição marginal, exceto quando a especificação utiliza uma ou duas distribuições marginais exponenciais, em que é contado um parâmetro para cada distribuição exponencial que há na especificação.

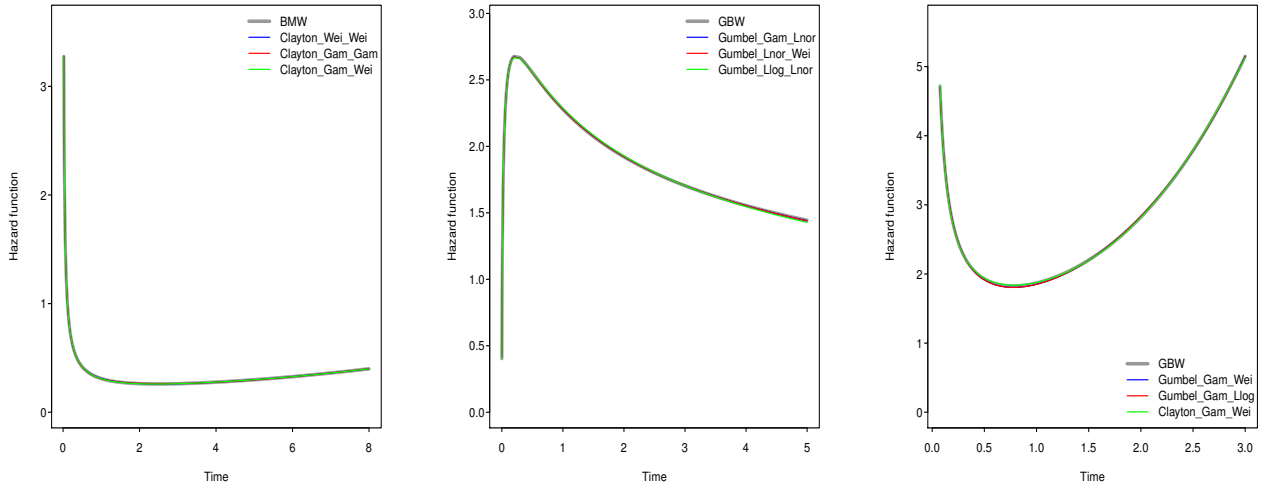


Figura 3.3: *Melhores ajustes de funções de taxa de falha beta modificada Weibull por funções de taxa de falha da família de riscos múltiplos. (a) distribuição BMW com parâmetros* $(a = 1,0;\ b = 0,7;\ \alpha = 1,0;\ \beta = 0,3;\ \mu = 10,0)$*, curva em preto, Figura 2 em Nadarajah et al. (2011); (b) distribuição GBW com* $(a = 2,0;\ b = 2,5;\ c = 1,2;\ \beta = 0,65;\ \mu = 0,5)$*, curva azul, Figura 7 em Alexander et al. (2011); (c) distribuição GBW com* $(a = 1,0;\ b = 1,5;\ c = 0,05;\ \beta = 3,0;\ \mu = 2,0)$*, curva verde, Figura 7 em Alexander et al. (2011).*

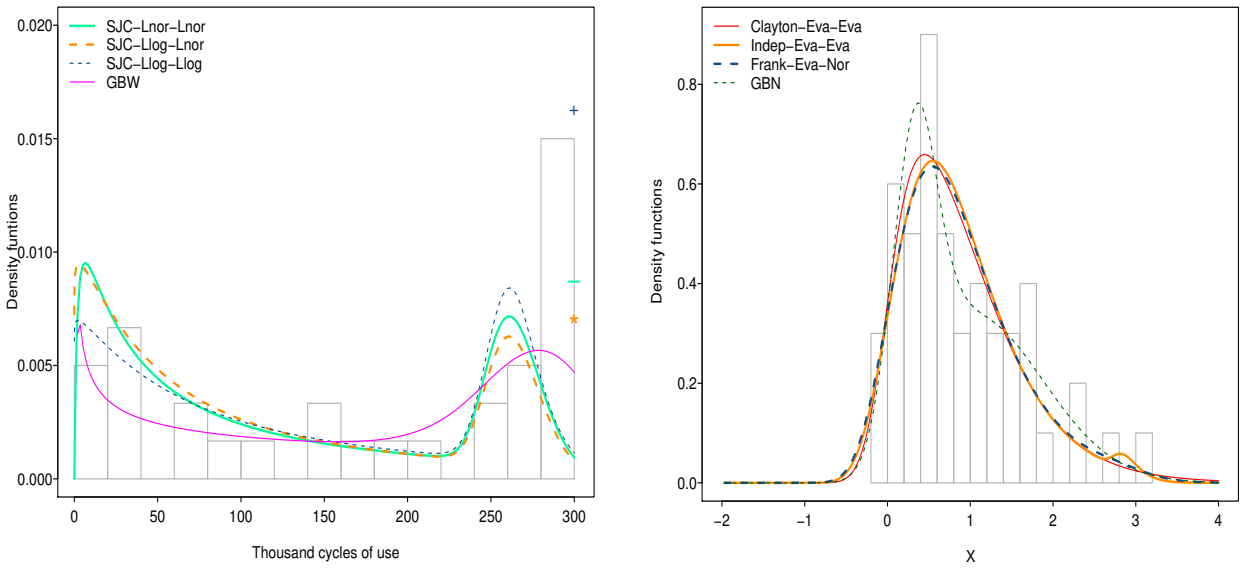


Figura 3.4: *(a) Funções densidade estimadas para os dados de voltagem por densidades PHD e pelo modelo GBW ajustado em Alexander et al. (2012). Há 8 observações censuradas no nível de voltagem igual a 300. O gráfico também mostra a função de sobrevivência estimada nesse valor. (b) Funções densidade estimadas para os dados da amostra skew normal por densidades PHD e pelo modelo GBN ajustado em Alexander et al. (2012). Em ambos os casos, são mostradas as três melhores densidades da família PHD selecionadas pelo critério AIC.*

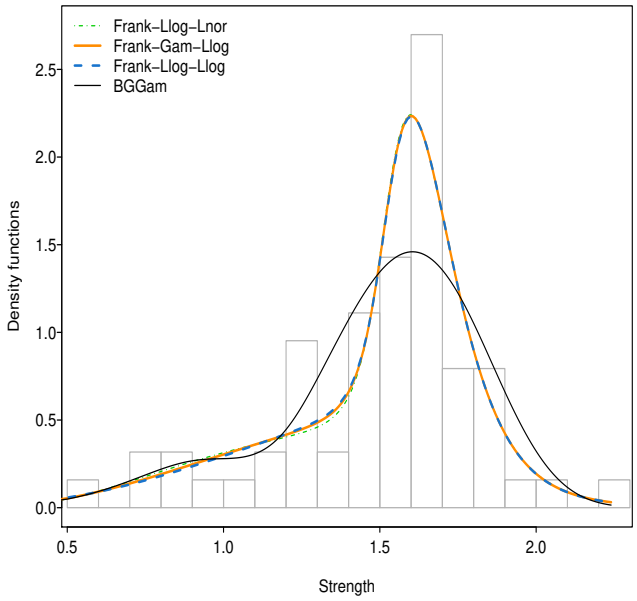


Figura 3.5: *Melhores ajustes por densidades do modelo PHD com cópula de Frank e ajuste pela densidade BGGam aos dados de fibras.*

Capítulo 4

Modelo de Misturas Generalizado Utilizando Cópulas

Nesse capítulo é proposta uma generalização do modelo de mistura de distribuições utilizando cópulas. Modelos de misturas finitas de distribuições são capazes de modelar uma ampla faixa de formas para representar conjuntos de dados com poucos parâmetros e por isso são bastante utilizados na literatura. Duas referências sobre esses modelos são os livros Titterington *et al.* (1985) e Mclachlan e Peel (2000). Segundo Mclachlan e Peel (2000), p.1, "os modelos de misturas finitas têm sido aplicados com sucesso em áreas como astronomia, biologia, genética, medicina, psiquiatria, economia, marketing e outros campos de ciências sociais e biológicas. Nessas aplicações os modelos de misturas finitas embasam várias técnicas de áreas da estatística como análise de agrupamento e estruturas latentes, análise discriminante, análise de imagens e análise de sobrevivência, além do papel que desempenham diretamente em análise de dados e inferência providenciando modelos descritivos para distribuições".

Algumas referências de modelos de misturas nas áreas de análise de sobrevivência e confiabilidade são Jewel (1982), que propõe um modelo de mistura de distribuições exponenciais; Jiang e Murthy (1998), que discutem formas possíveis para a função taxa de falha em modelos de mistura de distribuições Weibull; Bucar *et al.* (2004), que analisam o modelo de misturas de distribuições Weibull em sistemas de confiabilidade; Louzada-Neto *et al.* (2002), que propõem o modelo de misturas de funções de taxa de falha; Marín *et al.* (2005) e Tsionas (2002) que consideram a estimação bayesiana do modelo de misturas de distribuições Weibull.

A forma tradicional de construir um modelo de misturas é determinar parâmetros de misturas para cada distribuição componente. Por exemplo, para k funções densidade componentes $f_i(t)$, $i = 1, \cdots, k$, parâmetros de mistura não negativos p_i, tais que $\sum_{i=1}^{k} p_i = 1$, são utilizados para produzir o modelo com função densidade $\tilde{f}(t) = \sum_{i=1}^{k} p_i f_i(t)$.

Nesse trabalho propomos uma alternativa de modelagem em que são selecionados para a mistura quantis das distribuições componentes que são correspondentes a probabilidades de eventos de variáveis aleatórias distribuídas segundo uma função de cópula bivariada. Dessa definição resultam famílias de modelos que são parametrizados pela dependência da cópula e também por probabilidades associadas à participação das distribuições componentes na mistura. As proporções p_i são, então, substituídas por funções peso $p_i(t)$, que não são restritas à soma unitária para todos os valores de t. Essa modelagem acrescenta uma maior flexibilidade aos modelos de misturas na representação de dados com densidades multimodais e contendo efeitos de assimetria e, no caso de se modelar tempos de vida, com funções densidade e de taxa de falha de múltiplas curvaturas. O modelo proposto tem como casos particulares o modelo de mistura tradicional, os modelos de riscos múltiplos dependente e independente, o modelo de fração de cura e também gera famílias de modelos capazes de selecionar para a mistura determinadas regiões do suporte das distribuições componentes.

Na Seção 4.1 o modelo de misturas generalizado utilizando cópulas é definido, suas propriedades são discutidas e são apresentadas algumas famílias de modelos derivadas e exemplos de modelos da literatura que são seus casos particulares. Na Seção 4.2 algumas famílias apresentadas na Seção 2 são utilizadas para ilustrar a modelagem de assimetria e multimodalidade e também a capacidade do modelo de ponderar determinadas regiões do suporte das distribuições componentes para participação na mistura. Na Seção 4.3 são apresentadas algumas formas de produzir misturas com mais de duas distribuições componentes usando a definição proposta. Na Seção 4.4 o modelo é aplicado a conjuntos de dados simulados e também a dados reais analisados com outros modelos na literatura. Os procedimentos de estimação são baseados no método de máxima verossimilhança, não fazendo parte dos objetivos da tese investigar métodos alternativos de estimação do modelo.

4.1 O Modelo de Misturas Generalizado Utilizando Cópulas

O modelo de misturas generalizado utilizando cópulas é definido como:

Definição 6 *Considere U e V variáveis aleatórias uniformes em $[0,1]$ tais que $(U,V) \sim C$, em que C é uma função de cópula. Considere também um conjunto $A \subset [0,1]^2$ e F e G funções de distribuição. O modelo de misturas generalizado utilizando cópulas é definido pela variável T dada por*

$$T = F^{-1}(r(U,V))I_{[(U,V)\in A]} + G^{-1}(s(U,V))I_{[(U,V)\in A^c]}, \tag{4.1}$$

em que as funções r e s são tais que $r(u,v) \in [0,1]$, para $(u,v) \in A$, $s(u,v) \in [0,1]$, para $(u,v) \in A^c$. Se $P[r(U,V) = 0] > 0$, então existe um a_0 tal que $F(a_0) = 0$. Restrições similares são feitas quando $P[r(U,V) = 1] > 0$ e para $s(U,V)$. ■

A partir da Definição 6, temos que a variável T seleciona o quantil de probabilidade $r(u,v)$ de F ou $s(u,v)$ de G, de acordo com a ocorrência de (U,V) em A ou A^c por meio da cópula C. Dessa forma, as funções F e G participam como distribuições componentes na mistura do modelo. Também, dessas componentes são tomados para a mistura os quantis correspondentes às funções $r(u,v)$ e $s(u,v)$, que são chamadas funções de seleção dos quantis. Essa seleção, para a componente F, é função dos pares (u,v) pertencentes ao conjunto A e, para G, é função dos pares (u,v) pertencentes ao conjunto A^c. Por esse motivo, as regiões A e $A^c \subset [0,1]^2$ são chamados de conjuntos da mistura. Portanto, o modelo é definido pela sêxtupla (F,G,C,A,r,s).

O modelo também pode ser interpretado como uma soma de duas transformações $F^{-1}(r(U,V))$ e $G^{-1}(s(U,V))$ de variáveis aleatórias dependentes U e V com a restrição de que $(U,V) \in A$ para a primeira transformação e $(U,V) \in A^c$ para a segunda transformação. Como A e A^c são mutuamente exclusivos, T é dada pela primeira transformação com probabilidade $P(A)$ e pela segunda transformação com probabilidade $P(A^c)$. Como mostrado a seguir, o uso de F e G como duas funções de distribuição leva à mistura de suas respectivas funções densidade.

Nas discussões a seguir consideraremos que as distribuições F e G, e a variável aleatória T são (absolutamente) contínuas. Uma observação a respeito da continuidade de T é que mesmo quando F, G, C, r e s são contínuas, T não necessariamente é contínua. E, no caso que F e/ou r são funções degrau e G e/ou s também são funções degrau, resulta que T é uma função discreta.

A função de distribuição do modelo de misturas generalizado utilizando cópulas é dada por

$$\begin{aligned} P[T \le t] &= P[F^{-1}(r(U,V))I_{[(U,V)\in A]} + G^{-1}(s(U,V))I_{[(U,V)\in A^c]} \le t] \\ &= P[r(U,V) \le F(t), (U,V) \in A] + P[s(U,V) \le G(t), (U,V) \in A^c] \\ &= \int\int_{[(u,v)\in A; r(u,v)\le F(t)]} dC(u,v) + \int\int_{[(u,v)\in A^c; s(u,v)\le G(t)]} dC(u,v) \end{aligned} \tag{4.2}$$

Para o conjunto da mistura A dado, a determinação de $r(u,v)$ e $s(u,v)$ para especificar o modelo pode fazer com que $P(T \leq t)$ tenha ou não pontos de descontinuidade. Essa análise é realizada em alguns exemplos do modelo apresentados mais adiante nesta seção.

A partir de (4.2), a função densidade do modelo é dada por

$$\begin{aligned}\tilde{f}(t) &= \frac{d}{dt}P[T \leq t] \\ &= \frac{d}{dx}\left[\iint_{[(u,v)\in A; r(u,v)\leq x]} dC(u,v)\right]_{x=F(t)} f(t) + \frac{d}{dx}\left[\iint_{[(u,v)\in A^c; s(u,v)\leq x]} dC(u,v)\right]_{x=G(t)} g(t), \quad (4.3)\end{aligned}$$

para $f(t)$ e $g(t)$ as funções densidade correspondentes às funções de distribuição $F(t)$ e $G(t)$. De (4.3) destacamos as funções peso de participação das densidades f e g na mistura que são, respectivamente,

$$m(t) = \frac{d}{dx}\left[\int\int_{[(u,v)\in A; r(u,v)\leq x]} dC(u,v)\right]_{x=F(t)} \quad (4.4)$$

e

$$n(t) = \frac{d}{dx}\left[\int\int_{[(u,v)\in A^c; s(u,v)\leq x]} dC(u,v)\right]_{x=G(t)}. \quad (4.5)$$

A função densidade $\tilde{f}(t)$, em (4.3), é uma mistura das densidades $f(t)$ e $g(t)$, em que os pesos $m(t)$ e $n(t)$ são dependentes de t. Esses pesos são não negativos e não necessariamente têm soma unitária. Para visualizar essa propriedade, denominamos os eventos X_t, indexados por t real, em que $X_{t_1} \subseteq X_{t_2}$, para $t_1 \leq t_2$, como monotonicamente crescentes em t. Como $D_t = [(u,v) \in A; r(u,v) \leq F(t)]$ e $E_t = [(u,v) \in A^c; s(u,v) \leq G(t)]$ são eventos monotonicamente crescentes em t, segue que as derivadas da probabilidade dos eventos D_t e E_t em relação a t serão não negativas, quando definidas. Como resultado, $m(t)$ e $n(t)$ em (4.3) são funções não negativas que ponderam diferentemente as regiões do suporte de $f(t)$ e $g(t)$, respectivamente, de acordo com a especificação do modelo. Essas funções peso são ilustradas para algumas especificações do modelo na Seção 4.2.

No caso em que as distribuições F e G são funções de distribuição de variáveis aleatórias positivas, podemos utilizar a distribuição final para modelar tempos de vida. Neste caso, calculamos a função taxa de falha do modelo (4.2) da forma tradicional por $\tilde{h}(t) = \tilde{f}(t)/\tilde{S}(t)$, em que $\tilde{S}$ é a função de sobrevivência de T.

A Definição 6 é aplicada à mistura de duas distribuições, mas a mistura de mais de duas distribuições pode ser estendida diretamente, como analisado na Seção 4.3.

Consideramos a seguir alguns exemplos de especificação do modelo de misturas que geram famílias de distribuições e também que incluem alguns modelos da literatura como seus casos particulares. Essas especificações do modelo são analisadas de acordo com os conjuntos da mistura A e $A^c \subset [0,1]^2$, as funções de seleção $r(u,v)$ e $s(u,v)$ utilizadas para a variável T e dependência positiva, negativa ou independência da cópula C. Entre os casos particulares estão o modelo de misturas tradicional, o modelo de riscos múltiplos com dependência e o modelo de fração de cura, e entre as familias de distribuições são apresentadas especificações destinadas a selecionar determinados trechos do suporte das distribuições componentes, como regiões à esquerda, à direita, próximas ao centro e das caudas, produzindo modelos multimodais e com efeitos de assimetria dados pela mistura. O primeiro exemplo se refere ao modelo de misturas tradicional.

Exemplo 1 O modelo de misturas tradicional.

O modelo de misturas de duas distribuições $F(t)$ e $G(t)$ ponderadas pelas constantes p e $1-p$ com

$p \in (0,1)$, tem função de distribuição dada por

$$P(T \leq t) \;=\; p\,F(t) + (1-p)\,G(t),$$

e o denominamos modelo de misturas tradicional. Uma especificação do modelo de misturas utilizando cópulas que inclui o modelo de misturas tradicional é dada pela definição da variável T como

$$T \;=\; F^{-1}\left(\frac{U}{p}\right) I_{[U\leq p]} + G^{-1}\left(\frac{1-U}{1-p}\right) I_{[U>p]}, \tag{4.6}$$

em que $A = [(u,v) \in [0,1]^2; u \leq p]$, $r(u,v) = u/p$, $s(u,v) = (1-u)/(1-p)$, e podemos utilizar qualquer cópula. Nesse caso, a função de distribuição é dada por

$$\begin{aligned} P(T \leq t) &= P[F^{-1}(U/p) \leq t | U \leq p]\, P(U \leq p) + P[G^{-1}((1-U)/(1-p)) \leq t | U > p]\, P(U > p) \\ &= p\, P[U/p \leq F(t) | U \leq p] + (1-p)\, P[(1-U)/(1-p) \leq G(t) | U > p] \\ &= p\, F(t) + (1-p)\, G(t), \end{aligned}$$

dado que as distribuições de U/p condicionada a $U \leq p$ e de $(1-U)/(1-p)$ condicionada a $U > p$ são uniformes em $(0,1)$. Temos, então, o modelo de misturas tradicional.

Note que se tivermos em (4.6) a especificação da função de seleção $r(u,v) = u/q$, com $q > p$, segue que o modelo é dado pela distribuição

$$P(T \leq t) \;=\; q\,F(t)\, I_{[t\leq F^{-1}(p/q)]} + p\, I_{[t>F^{-1}(p/q)]} + (1-p)\,G(t),$$

que tem um ponto de descontinuidade em t igual ao quantil de probabilidade p/q de F. Também, se $q < p$, segue que $r(u,v)$ não é limitada em $[0,1]$. ■

Exemplo 2 O modelo de fração de cura.

Os modelos de sobrevivência com fração de cura, vide Capítulo 2, podem ser vistos como uma mistura tradicional de distribuições em que uma das funções de distribuição, que modela a parcela curada da população, vale zero em todo o seu suporte, exceto no tempo tendente ao infinito, quando assume valor um. Dessa forma, esses modelos são um caso particular do modelo de misturas generalizado. Considere a variável T definida por

$$T = F^{-1}(V) I_{[U\leq p]} + G^{-1}(V) I_{[U>p]} \tag{4.7}$$

em que $A = [(u,v), u \leq p]$, F é uma função tal que $F(x) = 0$ se $x < \infty$ e G é uma função de distribuição. Nesse caso

$$\begin{aligned} P[T \leq t] &= P[V \leq F(t), U \leq p] + P[V \leq G(t), U > p] \\ &= C(p, F(t)) + G(t) - C(p, G(t)) \\ &= G(t) - C(p, G(t)), \end{aligned} \tag{4.8}$$

em que a última igualdade considera $t < \infty$. Se é utilizada a cópula independência, então segue que $P[T \leq t] = (1-p)G(t)$, do que temos

$$\begin{aligned} P[T > t] &= 1 - (1-p)G(t) \\ &= p + (1-p)\bar{G}(t), \end{aligned} \tag{4.9}$$

em que $\bar{G}(t) = 1 - G(t)$. Para $\bar{G}$ a função de sobrevivência da parcela não curada da população, temos o modelo de fração de cura. Outras formas para s podem levar ao modelo de sobrevivência com fração

de cura, flexibilizando G em (4.8), como por exemplo $s(u,v) = v^2$. O uso de outras cópulas permite a fração de cura variar no tempo, de acordo com a função peso. As Figuras 4.5 e 4.6, por exemplo, ilustram a variação das funções peso.

Mazucheli et al. (2012) propõem para $\bar{G}$ o uso de modelos riscos múltiplos independentes. A versão com dependência desse modelo como um modelo de misturas é discutida no exemplo a seguir.

Exemplo 3 O modelo de riscos múltiplos.

O modelo de riscos múltiplos com dependência é discutido no Capítulo 3 e no Apêndice A. Considerando duas causas de falha latentes, o modelo tem função de sobrevivência dada por

$$S(t;\Upsilon) \;=\; C_\theta(S_1(t;\Gamma_1), S_2(t;\Gamma_2)), \tag{4.10}$$

em que C_θ é uma função de cópula com parâmetro de dependência desconhecido θ, $S_1(t;\Gamma_1)$ e $S_2(t;\Gamma_2)$ são as funções de sobrevivência dos tempos latentes de falha com parâmetros desconhecidos Γ_1 e Γ_2, e $\Upsilon = (\theta, \Gamma_1, \Gamma_2)$. Esse modelo resulta da função de sobrevivência conjunta dos tempos latentes, dada por

$$S(t_1, t_2; \Upsilon) \;=\; C_\theta(S_1(t_1;\Gamma_1), S_2(t_2;\Gamma_2)). \tag{4.11}$$

Para a análise do modelo de misturas generalizado incluindo o modelo de riscos múltiplos com dependência, defina T como

$$T = F^{-1}(U) I_{[F^{-1}(U) \le G^{-1}(V)]} + G^{-1}(V) I_{[F^{-1}(U) > G^{-1}(V)]}, \tag{4.12}$$

em que $A = [U \le F \circ G^{-1}(V)]$, $r(u,v) = u$, $s(u,v) = v$ e podemos usar qualquer cópula, que denotamos por C.

Podemos escrever a variável aleatória T como $T = \min(T_1, T_2)$, em que $T_1 = F^{-1}(U)$ e $T_2 = G^{-1}(V)$ são variáveis aleatórias com distribuições F e G, respectivamente. Os eventos $[T_1 = \min(T_1, T_2)]$ e $[T_2 = \min(T_1, T_2)]$ formam os conjuntos da mistura A e A^c, respectivamente, mostrados em (4.12). Então temos o modelo de riscos múltiplos com dependência escrito em termos de funções de distribuição. Para verificar a equivalência desse modelo com (4.10), basta verificar a equivalência das distribuições conjuntas. Note que a distribuição conjunta de T_1 e T_2, que denotamos por H, é verificada por $U = F(T_1)$ e $V = G(T_2)$ e $(U,V) \sim C$, do que segue que $H(t_1,t_2) = C(F(t_1), G(t_2))$. Fazendo $F(t) = 1 - S_1(t;\Gamma_1)$, $G(t) = 1 - S_2(t;\Gamma_2)$ e C dada pela cópula de sobrevivência correspondente a C_θ, isto é, $C(u,v) = u + v - 1 + C_\theta(1-u, 1-v)$, temos

$$\begin{aligned} H(t_1,t_2) &= C(F(t_1), G(t_2)) \\ &= 1 - S_1(t_1;\Gamma_1) - S_2(t_2;\Gamma_2) + C_\theta(S_1(t_1;\Gamma_1), S_2(t_2;\Gamma_2)), \end{aligned} \tag{4.13}$$

que é a função de distribuição conjunta correspondente à função de sobrevivência (4.11).

Quando é utilizada a cópula independência temos o modelo de riscos múltiplos tradicional e temos o modelo de riscos múltiplos dependente quando utilizamos uma cópula não independente.

Essa forma generalizada de escrever o modelo de riscos múltiplos permite construir outros modelos a partir da alteração das funções utilizadas na sua especificação. Para um exemplo, fazendo $s(u,v) = 1-v$, o modelo passa a tomar para a mistura quantis de G opostos em relação à mediana, levando a um modelo de concorrência do primeiro tempo em ser um valor de mínimo mais extremo que o segundo tempo ser um valor máximo extremo, em termos das probabilidades dos quantis das distribuições componentes. Outra forma de modificar o modelo é definir A e A^c de forma trocada, isto é, tal que $A = [F^{-1}(U) > G^{-1}(V)]$. Nesse caso o modelo de misturas equivale à concorrência entre os tempos latentes em ser o máximo valor realizado. ■

Exemplo 4 Modelos de mistura generalizados usando conjuntos simples (MGC-CS).

Os modelos deste exemplo são denominados modelos de mistura por conjuntos simples, devido à formação dos conjuntos de mistura A e A^c, dados pela separação, em relação à uma constante, de uma das dimensões entre u e v e de forma independente da outra. Nessa família, aos conjuntos de misturas de forma definida no Exemplo 1, dada por $A = [(u, v) \in [0, 1]^2; u \leq p]$, são aplicadas outras funções de seleção $r(u, v)$ e $s(u, v)$ resultando em uma família de modelos capaz de misturar regiões à esquerda ou à direita das distribuições componentes. Essas regiões são tomadas de forma oposta para cada componente e, conforme a dependência da cópula, são localizadas mais próximas às caudas ou ao centro. O número de parâmetros dos modelos é dado pelo número de parâmetros da cópula, das distribuições componentes F e G e da variável da mistura, p. Essa família também inclui o modelo de misturas tradicional e o modelo de fração de cura, quando a cópula equivale à independência.

Considere a especificação dada por $A = [(u, v) \in [0, 1]^2; u \leq p]$ e $r(u, v) = s(u, v) = v$. Segue que a variável T, dada por

$$T = F^{-1}(V)I_{[U \leq p]} + G^{-1}(V)I_{[U>p]}, \tag{4.14}$$

tem função de distribuição

$$\begin{aligned} P(T \leq t) &= P[F^{-1}(V) \leq t, U \leq p] + P[G^{-1}(V) \leq t, U > p] \\ &= P[V \leq F(t), U \leq p] + P[V \leq G(t), U > p] \\ &= C(p, F(t)) + G(t) - C(p, G(t)), \end{aligned} \tag{4.15}$$

e função densidade

$$\tilde{f}(t) = C_2'(p, F(t))\, f(t) + \left[1 - C_2'(p, G(t))\right] g(t), \tag{4.16}$$

em que $C_2'(a, b) = \frac{d}{dx} C(a, x)|_{x=b}$.

Nesse modelo, o parâmetro p determina a ponderação dada à primeira parcela de T em (4.14), que é a transformação dada pela função composta $F^{-1} \circ r$ de (U, V), restando a ponderação de $(1-p)$ para a segunda parcela, $G^{-1} \circ s$. Para visualizar a mistura em termos das distribuições F e G, se a dependência da cópula é negativa, então são tomados para a mistura os quantis de alta probabilidade de F e de baixa probabilidade de G. O resultado é invertido para as regiões do suporte de F e G tomadas para a mistura se a dependência é positiva. Se é utilizada a cópula independência no modelo, então de (4.15) temos que

$$\begin{aligned} P(T \leq t) &= C(p, F(t)) + G(t) - C(p, G(t)) \\ &= p\, F(t) + (1 - p)\, G(t) \end{aligned} \tag{4.17}$$

o que é o modelo de misturas tradicional. Essa família é denominada modelos de mistura de valores de regiões opostas das distribuições usando conjuntos simples e utilizamos a sigla MGC-CS-O. Essa família é analisada no Exemplo 11 e utilizada para ajustar os dados de acidez da água em lagos, na Seção 4.4.2, e dos tempos de erupção do gêiser The Old Faithful, na Seção 4.4.2.

Uma outra família de modelos MGC-CS é obtida se forem tomadas $r(u, v) = u/p$ e $s(u, v) = v$. O modelo resultante é dado pela função de distribuição

$$\begin{aligned} P(T \leq t) &= P[F^{-1}(U/p) \leq t, U \leq p] + P[G^{-1}(V) \leq t, U > p] \\ &= P[U \leq pF(t), U \leq p] + P[V \leq G(t), U > p] \\ &= p\, F(t) + G(t) - C(p, G(t)), \end{aligned} \tag{4.18}$$

e função densidade

$$\tilde{f}(t) \quad = \quad p\, f(t) + \left[1 - C_2'(p, G(t))\right] g(t). \tag{4.19}$$

Nessa especificação, o modelo permite, de acordo com a dependência da cópula, a tomada de valores das regiões à esquerda ou à direita apenas da distribuição G, enquanto que da distribuição F é tomada toda a distribuição. Nesse caso temos a família de modelos de mistura de valores de regiões à esquerda ou à direita de uma das componentes usando conjuntos simples e usamos a sigla MGC-CS-CED, pois a primeira componente participa de forma completa, sem seleção de regiões. Essa família é utilizada para ajustar aos dados de atividade enzimática, na Seção 4.4.2. Tomando-se $G = F$, o modelo possibilita a inclusão de efeitos de assimetria à esquerda e à direita da distribuição F, enquanto que, também, tem menos parâmetros. Nesse caso, é atribuída à distribuição original F a probabilidade p e ao efeito de assimetria, $G^{-1}(s)$, a probabilidade $1 - p$. O efeito de assimetria varia em intensidade e tipo (à esquerda ou à direita) de acordo com a dependência da cópula. Essa família é denominada por modelos de mistura de efeito de assimetria usando conjuntos simples (MGC-CS-EA) - vide Exemplo 9.

O modelo também inclui o modelo de misturas tradicional como caso particular quando a cópula é independência. Como no Exemplo 1, a especificação de $q = p$ em $r(u, v) = u/q$, se deve à continuidade da função de distribuição de $P(T \leq t)$.

Na família de conjuntos simples MGC-CS-O ocorre um tipo de não identificabilidade. Para duas especificações A e B, sob as condições de mesma cópula simétrica, $\tau_A = -\tau_B$ e $p_A = 1 - p_B$, segue que a ordem das componentes F e G é importante para a identificabilidade do modelo. Nesse caso, as duas especificações selecionam os mesmos valores das componentes se estas forem trocadas de posição ($F_A = G_B$ e $F_B = G_A$), e, portanto, resultam no mesmo modelo de misturas. Esse tipo de não identificabilidade não ocorre nas famílias MGC-CS-CED e MGC-CS-EA, pois a tomada de valores da primeira e da segunda componente para a mistura ocorre de formas distintas, em que de uma é tomada a faixa completa do suporte e de outra, prioritariamente as regiões à esquerda ou à direita do seu suporte, conforme a dependência da cópula. ■

Exemplo 5 Modelos de mistura generalizados usando conjuntos de reflexão (MGC-CR)

Nesse exemplo a especificação do modelo utiliza os conjuntos da mistura dados por A=$[(u, v); u \leq v]$. Conforme a determinação das funções de seleção, $r(u, v)$ e $s(u, v)$, é possível construir modelos que selecionam para a mistura valores de diversos trechos das distribuições componentes entre regiões à esquerda de ambas distribuições, regiões à esquerda de uma distribuição e à direita de outra, regiões das caudas de ambas distribuições etc. Os modelos dessa família são denominados modelos de mistura usando conjuntos de reflexão devido à simetria dos conjuntos de mistura A e A^c em relação à linha $u = v$. O número de parâmetros dos modelos é dado pelo número de parâmetros da cópula e das distribuições componentes F e G. Também, nessa família temos $P[A]$ e $P[A^c]$ para os valores tomados de cada transformação $F^{-1} \circ r$ e de $G^{-1} \circ s$. Nesse trabalho uma cópula simétrica $C(u, v)$ é aquela em que seus argumentos u e v são intercambiáveis, isto é, $C(u, v) = C(v, u)$ para qualquer (u, v) em $[0, 1]^2$. Para os modelos das famílias a seguir são utilizadas cópulas simétricas.

Considere o modelo de misturas dado pela variável T, definida por

$$T = F^{-1}(r(U, V))\, I_{[U \leq V]} + G^{-1}(s(U, V))\, I_{[V < U]}, \tag{4.20}$$

em que $(U, V) \sim C$ e C é uma cópula simétrica. As regiões A e A^c são definidas pela função $v = u$ em $[0, 1]^2$, o que equivale a se tomar o mínimo observado entre U e V para o cálculo do quantil de F ou de G, conforme (4.20). As probabilidades $r(u, v)$ e $s(u, v)$ para a seleção dos quantis de F e G podem definir misturas de diferentes regiões do suporte dessas distribuições componentes e consideramos

1. *$r_a(u, v) = u$ e $s_a(u, v) = v$;*

2. $r_b(u,v) = 1-u$ *e* $s_b(u,v) = v$;

3. $r_c(u,v) = [\chi(2u-1)|2u-1|^{1/k_1}+1]/2$ *e* $s_c(u,v) = [\chi(2v-1)|2v-1|^{1/k_2}+1]/2$

para k_i, $i = 1,2$, *parâmetros reais tais que* $k_i > 1$, *e* $\chi(x)$, *a função sinal de* x. *Com essas especificações de funções de seleção os modelos são chamados de modelos de mistura de valores de regiões à esquerda das distribuições (MGC-CR-EE), de mistura de valores de regiões à esquerda e à direita das distribuições (MGC-CR-ED) e de mistura de valores das caudas das distribuições com assimetria à esquerda usando conjuntos de reflexão (MGC-CR-CAE). A família MGC-CR-EE é analisada nos Exemplos 7 e 10 e ilustra o ajuste a dados simulados na Seção 4.4.1 (Caso 3). O Exemplo 8 analisa a família MGC-CR-ED. A família MGC-CR-CAE é analisada no Exemplo 12 e utilizada na Seção 4.4.1 na análise de ajuste a dados simulados (Caso 2). Modelos de mistura de valores de regiões à direita das distribuições (MGC-CR-DD) podem ser construídos por* $r(u,v) = 1-u$ *e* $s(u,v) = 1-v$ *e modelos de mistura de valores das caudas das distribuições com assimetria à direita (MGC-CR-CAD), por* $r(u,v) = [1-\chi(2u-1)|2u-1|^{1/k_1}]/2$ *e* $s(u,v) = [1-\chi(2v-1)|2v-1|^{1/k_2}]/2$.

Segundo (4.2), a distribuição de T *é dada por*

$$\begin{aligned} P(T\le t) &= P[F^{-1}(r(U,V))\, I_{[U\le V]} + G^{-1}(s(U,V))\, I_{[V<U]} \le t] \\ &= P[r(U,V)\le F(t), U\le V] + P[s(U,V)\le G(t), V<U] \\ &= \int\int_{[(u,v)\in[0,1]^2; U\le V; r(u,v)\le F(t)]} dC(u,v) + \int\int_{[(u,v)\in[0,1]^2; U>V; s(u,v)\le G(t)]} dC(u,v) \quad (4.21) \end{aligned}$$

para $c(x,y)$ *a função densidade da cópula* C. *A cópula simétrica e as regiões* A *e* A^c *dadas por* $A = [U\le V]$ *e* $A^c = [V<U]$ *facilitam que* $P(T\le t)$ *seja escrita diretamente em função da cópula* C, *simplificando o cálculo integral e a forma analítica da função de distribuição (4.21). Nesse caso, as funções de distribuição correspondentes às funções de seleção* r_i *e* s_i, $i\in\{a,b,c\}$, *são simplificadas para*

$$P_a(T\le t) = F(t) - \frac{\delta_C(F(t))}{2} + G(t) - \frac{\delta_C(G(t))}{2} \quad (4.22)$$

$$P_b(T\le t) = F(t) - \frac{1}{2} + \frac{\delta_C(1-F(t))}{2} + G(t) - \frac{\delta_C(G(t))}{2} \quad (4.23)$$

$$\begin{aligned} P_c(T\le t) = 1 &+ \frac{\chi(2F(t)-1)|2F(t)-1|^{k_1}}{2} - \frac{\delta_C(\frac{\chi(2F(t)-1)|2F(t)-1|^{k_1}+1}{2})}{2} \\ &+ \frac{\chi(2G(t)-1)|2G(t)-1|^{k_2}}{2} - \frac{\delta_C(\frac{\chi(2G(t)-1)|2G(t)-1|^{k_2}+1}{2})}{2}, \quad (4.24) \end{aligned}$$

em que $\delta_C(x) = C(x,x)$, *a diagonal da cópula. As funções densidade correspondentes a (4.22)-(4.24) são dadas por*

$$\tilde f_a(t) = \left[1-\delta_C'(F(t))/2\right] f(t) + \left[1-\delta_C'(G(t))/2\right] g(t) \quad (4.25)$$

$$\tilde f_b(t) = \left[1+\delta_C'(1-F(t))/2\right] f(t) + \left[1-\delta_C'(G(t))/2\right] g(t) \quad (4.26)$$

$$\begin{aligned} \tilde f_c(t) &= k_1|2F(t)-1|^{k_1-1}\left[1-\delta_C'(\frac{\chi(2F(t)-1)|2F(t)-1|^{k_1}+1}{2})/2\right] f(t) \\ &\quad + k_2|2G(t)-1|^{k_2-1}\left[1-\delta_C'(\frac{\chi(2G(t)-1)|2G(t)-1|^{k_2}+1}{2})/2\right] g(t), \quad (4.27) \end{aligned}$$

em que $\delta_C'(x) = \frac{d}{dx}\delta_C(x)$.

A esperança das variáveis da mistura correspondentes a (4.22) a (4.24) quando a cópula equivale à

independência são dadas, respectivamente, por

$$\begin{aligned} E[T_a] &= \mu_F - \frac{\mu_{F^2}}{2} + \mu_G - \frac{\mu_{G^2}}{2}, \\ E[T_b] &= \frac{\mu_{F^2}}{2} + \mu_G - \frac{\mu_{G^2}}{2}\, e \\ E[T_c] &= \mu_{\hat{F}(t)} - \frac{1}{2}\mu_{\hat{F}(t)^2} + \mu_{\hat{G}(t)} - \frac{1}{2}\mu_{\hat{G}(t)^2}, \end{aligned}$$

em que $\mu_{F^*} = E[X]$ *para* $X \sim F^*$, $\hat{X} = \chi(2X-1)|2X-1|^{k_1}+1]/2$ *e as funções* $F(t)^2$ *e* $[\chi(2F(t)-1)|2F(t)-1|^{k_1}+1]/2$ *são funções de distribuição.*

No primeiro caso, definido por r_a *e* s_a*, se a cópula tem dependência negativa, a mistura é composta de valores das regiões à esquerda de cada distribuição componente. Conforme a dependência da cópula aumenta, as regiões do suporte de* F *e* G *tomadas para a mistura são estendidas até que compreendam todo o suporte das componentes quando a dependência é máxima, o que resulta no modelo de mistura tradicional* $f(t)/2+g(t)/2$*. O segundo caso, definido por* r_b *e* s_b*, tem comportamento semelhante, porém, quando a dependência é negativa, são tomados valores da região à direita de* F *e da esquerda de* G*. Da mesma forma, conforme aumenta a dependência da cópula, essas regiões tendem a compreender todo o suporte das distribuições componentes, resultando no modelo de misturas tradicional particular* $f(t)/2 + g(t)/2$*. A terceira especificação, resultante das funções de seleção* r_c *e* s_c*, pode tomar para a mistura valores das caudas de* F *e de* G *atribuindo maior peso às caudas à esquerda quando a dependência é negativa até se equilibrar conforme a dependência aumenta até ter valor máximo.* ■

Uma observação é que se o conjunto A dado no Exemplo 5 é generalizado para $A = [U \leq aV]$, para $a \in (0,1)$, isso leva à definição da variável da mistura

$$T = F^{-1}(U)\, I_{[U/a \leq V]} + G^{-1}(s(U,V))\, I_{[V<U/a]}, \tag{4.28}$$

e função de distribuição

$$\begin{aligned} P[T \leq t] &= P[U \leq aF(t), U \leq aV] + P[V \leq G(t), U > aV] \\ &= \int_0^{aF(t)} \int_{u/a}^1 c(u,v)dvdu + \int_0^{G(t)} \int_0^a c(u,v)dudv + G(t) - C(a, G(t)). \end{aligned} \tag{4.29}$$

Essa especificação de conjuntos de mistura A e A^c pode proporcionar maior liberdade para selecionar os elementos (u,v) para tomar valores de F ou G, porém o modelo não pode ser simplificado quando a cópula é simétrica, restando o trabalho analítico do cálculo integral em (4.29).

No exemplo a seguir, a especificação dos conjuntos de mistura seguem $[0,1]^2$ particionado em três regiões, possibilitando relacionar duas dessas regiões a uma das distribuições componentes, facilitando a seleção dos valores desejados para participar da mistura. A família também é chamada de modelos de misturas por conjuntos simples pela definição dos conjuntos da mistura. O modelo também inclui o modelo de misturas tradicional quando a cópula equivale à independência.

Exemplo 6 Modelos de mistura generalizado usando três conjuntos simples (MGC-3CS).
Considere a especificação dada por $A = [(u,v) \in [0,1]^2; u \leq p \cup u > q]$ *para* $0 < p < q < 1$ *e* $r(u,v) = s(u,v) = v$*. Segue que a variável* T *dada por*

$$T = F^{-1}(V) I_{[U \leq p] \cup [U > q]} + G^{-1}(V) I_{[p < U \leq q]}, \tag{4.30}$$

tem função de distribuição

$$\begin{aligned} P(T \leq t) &= P[F^{-1}(V) \leq t, U \leq p] + P[F^{-1}(V) \leq t, U > q] + P[G^{-1}(V) \leq t, p < U \leq q] \\ &= P[V \leq F(t), U \leq p] + P[V \leq F(t), U > q] + P[V \leq G(t), p < U \leq q] \\ &= C(p, F(t)) + F(t) - C(q, F(t)) + C(q, G(t)) - C(p, G(t)), \end{aligned} \tag{4.31}$$

e função densidade

$$\tilde{f}(t) = \left[1 + C_2'(p, F(t)) - C_2'(q, F(t))\right] f(t) + \left[C_2'(q, G(t)) - C_2'(p, g(t))\right] g(t). \tag{4.32}$$

Nessa família o número de parâmetros dos modelos é dado pelo número de parâmetros da cópula, das distribuições componentes F e G e da variável da mistura, p e q. No modelo, se a dependência da cópula é alta e positiva, então são tomados para a mistura os quantis de F de probabilidade aproximadamente até p e após q e de G, aproximadamente entre p e q. Essa situação se inverte para a dependência alta negativa, em que participam da mistura os quantis de F de probabilidade aproximadamente até $1-q$ e após $1-p$ e de G, aproximadamente entre $1-q$ e $1-p$. Esse grau de aproximação é mais definido quanto maior é a dependência em módulo. Também, se é utilizada a cópula independência no modelo, então (4.31) equivale a $(1-q+p)F(t)+(q-p)G(t)$ que é o modelo de misturas tradicional. Essa família é chamada de modelos de mistura de valores das caudas e centro das distribuições usando conjuntos simples (MGC-3CS-CC). Essa família é analisada no Exemplo 13 e utilizada nos procedimentos de simulação na Seção 4.4.1 (Caso 1).

Uma particularização desse modelo é obtida fazendo $q = 1 - p$, para $0 < p < 1/2$, o que reduz o número de parâmetros e torna simétricas, em relação a $1/2$, as probabilidades dos quantis tomados das caudas de F e, portanto, também da região central de G. Nesse caso, a família é chamada de modelos de mistura de valores das caudas e centro das distribuições usando conjuntos simples com tomada simétrica dos quantis (MGC-3CS-CC-S).

Nas famílias MGC-3CS-CC e MGC-3CS-CC-S ocorre um tipo de não identificação quando a cópula é simétrica. Na família MGC-3CS-CC, para duas parametrizações A e B com mesmos parâmetros de cópula e distribuições componentes, se $\tau_A = -\tau_B$, $p_A = 1-q_B$ e $q_A = 1-p_B$, então A e B correspondem ao mesmo modelo. Para a família MGC-3CS-CC-S, as parametrizações A e B de mesmos parâmetros de cópula, distribuições componentes e parâmetro p correspondem ao mesmo modelo se $\tau_A = -\tau_B$. ■

4.2 Exemplos de Funções Densidade

Essa seção contém alguns exemplos de distribuições geradas pelo modelo proposto nesse capítulo. Apesar de o modelo utilizar quaisquer distribuições F e G, usamos nos exemplos a seguir o mesmo par de distribuições de forma a poder mostrar o potencial do método. As distribuições F e G são Weibull, com parâmetros $(\mu_1; \beta_1) = (2,0; 3,0)$ e $(\mu_2; \beta_2) = (7,0; 6,0)$, respectivamente. Também, é utilizada nas ilustrações a função de cópula de Frank devido à sua capacidade de representar toda a faixa de dependência negativa a positiva, o que permite que uma dada especificação do modelo represente mais formas em comparação com a situação de se utilizar cópulas que modelam apenas uma parte dessa faixa de dependência. Por serem utilizadas distribuições de suporte nos reais positivos é possível usar o modelo resultante para modelar tempos de vida. Em cada exemplo são mostradas as funções densidade, as funções peso e a função taxa de falha. Para a cópula, mostramos os resultados usando apenas três casos: o caso independente e quando o coeficiente τ de Kendall é igual a $-0,8$ e $0,8$.

Os dois exemplos a seguir ilustram a família de modelos MGC-CR.

Exemplo 7 *Considere a especificação do modelo de MGC-CR-EE, do Exemplo 5 e com as componentes Weibull e cópula de Frank.*

Nesse caso, a função de distribuição (4.22) é dada por

$$P(T \leq t) = F(t) + \frac{1}{2\theta}\log\left\{1 - \frac{(1-e^{-\theta F(t)})^2}{(1-e^{-\theta})}\right\} + G(t) + \frac{1}{2\theta}\log\left\{1 - \frac{(1-e^{-\theta G(t)})^2}{(1-e^{-\theta})}\right\} \quad (4.33)$$

e a função densidade, por

$$\tilde{f}(t) = \left\{1 - \frac{e^{-\theta F(t)} - e^{-2\theta F(t)}}{2e^{-\theta(F(t))} - e^{-\theta} - e^{2\theta F(t)}}\right\} f(t) + \left\{1 - \frac{e^{-\theta G(t)} - e^{-2\theta G(t)}}{2e^{-\theta(G(t))} - e^{-\theta} - e^{2\theta G(t)}}\right\} g(t). \quad (4.34)$$

A Figura 4.1 mostra nos gráficos (a), (b) e (c) as funções densidade componentes da mistura, as funções peso correspondentes às distribuições componentes Weibull e o modelo de misturas resultante para os casos de independência e de dependência dada pelo τ de Kendall igual a $-0,8$ e $0,8$. O mecanismo de dependência negativo da cópula resulta na mistura com maior ponderação para quantis de baixa probabilidade de cada uma das distribuições componentes, aproximadamente até às suas medianas. A dependência positiva alta e a simetria da cópula levam à seleção de observações de toda a faixa de valores possíveis das distribuições componentes com funções peso tendentes à função constante $p = 1/2$ para cada uma das duas distribuições conforme o coeficiente τ de Kendall se aproxima de 1.

A Figura 4.2 mostra nos nos gráficos (a), (b) e (c) as funções densidade, taxa de falha e sobrevivência do modelo resultante para os respectivos casos de dependência. O modelo exibe flexibilidade na mistura das distribuições em função do parâmetro de dependência da cópula. ■

Exemplo 8 *Considere a especificação do modelo MGC-CR-ED, dada no Exemplo 5, com as componentes Weibull e cópula de Frank. Nesse caso, temos que a função de distribuição (4.23) é dada por*

$$P(T \leq t) = F(t) - \frac{1}{2} - \frac{1}{2\theta}\log\left\{1 - \frac{(1-e^{-\theta\tilde{F}(t)})^2}{(1-e^{-\theta})}\right\} + G(t) + \frac{1}{2\theta}\log\left\{1 - \frac{(1-e^{-\theta G(t)})^2}{(1-e^{-\theta})}\right\} \quad (4.35)$$

e a função densidade, por

$$\tilde{f}(t) = \left\{1 + \frac{e^{-\theta\tilde{F}(t)} - e^{-2\theta\tilde{F}(t)}}{2e^{-\theta\tilde{F}(t)} - e^{-\theta} - e^{2\theta\tilde{F}(t)}}\right\} f(t) + \left\{1 - \frac{e^{-\theta G(t)} - e^{-2\theta G(t)}}{2e^{-\theta(G(t))} - e^{-\theta} - e^{2\theta G(t)}}\right\} g(t), \quad (4.36)$$

em que $\tilde{F}(t) = 1 - F(t)$.

A Figura 4.1 mostra nos gráficos (d), (e) e (f) as funções densidade componentes da mistura, as funções peso correspondentes às distribuições componentes Weibull e o modelo de misturas resultante. O mecanismo de dependência negativo da cópula leva à mistura com maior ponderação para os quantis de baixa probabilidade de uma das distribuições componentes e de alta probabilidade da outra. Ocorre o mesmo efeito de tendência à mistura tradicional de peso 1/2 para cada componente em relação ao Exemplo 7. A Figura 4.2 mostra nos gráficos (d), (e) e (f) as funções densidade, taxa de falha e sobrevivência do modelo resultante para os respectivos casos de dependência. As Figuras 4.1 e 4.2 mostram em (g), (h) e (i) as funções para o modelo resultante no caso em que são trocadas de posição no modelo as funções F e G. ■

Os dois exemplos a seguir ilustram o modelo de mistura de efeito de assimetria MGC-CS-EA e o modelo de MGC-CR-EE utilizando a mesma distribuição para F e G, com o objetivo de produzir um modelo assimétrico.

Exemplo 9 *As Figuras 4.3 e 4.4 mostram nos gráficos (a), (b) e (c), as formas obtidas se consideramos no modelo MGC-CS-EA, do Exemplo 4, em (4.18), ambas distribuições F e G dadas por uma*

distribuição, $Weibull(3,0;2,0)$, cópula de Frank e constante $p = 1/3$. O resultado é um modelo de misturas que atribui ponderação constante igual a p a todo o suporte de F e também atribui maior ponderação às regiões à esquerda de G se a dependência da cópula é negativa e à direita de G se a dependência da cópula é positiva. Dessa forma é construído um modelo com assimetria à esquerda ou à direita, parametrizado pela dependência da cópula. A assimetria é inexistente quando $\tau = 0$, em que o modelo equivale ao modelo de misturas tradicional de pesos constantes p e $1-p$ para F e G que, nesse caso, são distribuições iguais. Também as ponderações da distribuição original e da transformação $F^{-1} \circ s$ para assimetria são parametrizadas por p e $1-p$. ■

Exemplo 10 *Os gráficos (d), (e) e (f) das Figuras 4.3 e 4.4 mostram o modelo MGC-CR-EE, do Exemplo 5 com ambas distribuições F e G dadas por uma distribuição, $Weibull(3,0;2,0)$. Conforme a dependência da cópula diminui, com τ de Kendall variando de 1 a -1, é dada maior ponderação às regiões à esquerda da distribuição Weibull por F e G produzindo um modelo assimétrico à esquerda parametrizado pela dependência da cópula. Essa deformação é mínima ou inexistente quando $\tau = 1$, em que o modelo de misturas tem função peso constante igual a $1/2$ para cada uma das distribuições, que são iguais nesse caso, ou máxima quando $\tau = -1$, em que o modelo equivale à soma do modelo em duas regiões à esquerda da mediana da mesma distribuição.* ■

O exemplo a seguir ilustra o modelo MGC-CS-O.

Exemplo 11 *A Figura 4.5 ilustra a generalização do modelo de misturas tradicional pelo modelo MGC-CS-O, em (4.15) do Exemplo 4 com as distribuições F e G Weibull dadas, cópula de Frank e constante $p = 1/3$. Conforme a dependência da cópula varia de negativa a postiva, as funções peso refletem a seleção de quantis de alta probabilidade a baixa probabilidade da distribuição F e, simultaneamente, de quantis de baixa probabilidade a alta probabilidade da distribuição G. O modelo resultante é o da mistura de valores das regiões à esquerda de F e à direita de G quando a dependência é negativa, o modelo de misturas tradicional de pesos constantes $p = 1/3$ para F e $1-p = 2/3$ para G quando não há dependência, e a mistura das regiões à direita de F e à esquerda G quando a dependência é positiva.*

■

Os dois exemplos a seguir ilustram o modelo de misturas utilizado para representar distribuições que são misturas de partes das regiões das caudas e central das distribuições componentes. O primeiro é o modelo de misturas de valores das caudas das distribuições MGC-CR-CAE, utilizando funções de seleção em forma de *"s"* para tomar quantis de probabilidades baixas ou altas das distribuições componentes. O segundo é o modelo de misturas de valores das caudas e centro das distribuições MGC-3CS-CC, fazendo uma partição de $[0,1]^2$ em três regiões que, combinadas com cópulas de alta dependência (positiva ou negativa), faz seleção das caudas de uma componente e região central da outra.

Exemplo 12 *Considere a especificação do modelo MGC-CR-CAE, do Exemplo 5 com funções de seleção $r(u,v) = [1+(2u-1)^{1/k_1}]/2$, $s(u,v) = [1+(2v-1)^{1/k_2}]/2$, para $k_1 = 2,5$ e $k_2 = 3,0$, componentes Weibull e cópula de Frank.*

Nesse caso, temos que a função de distribuição (4.24) é dada por

$$\begin{aligned} P(T \leq t) \;=\; & 1 + \frac{(2F(t)-1)^{2,5}}{2} + \frac{1}{2\theta}\log\left\{1 - \frac{(1-e^{-\theta\frac{(2F(t)-1)^{2,5}+1}{2}})^2}{(1-e^{-\theta})}\right\} \\ & + \frac{(2G(t)-1)^3}{2} + \frac{1}{2\theta}\log\left\{1 - \frac{(1-e^{-\theta\frac{(2G(t)-1)^3+1}{2}})^2}{(1-e^{-\theta})}\right\} \end{aligned} \tag{4.37}$$

e a função densidade, por

$$\tilde{f}(t)=2,5(2F(t)-1)^{1,5}\left(1-\frac{e^{-\frac{\theta}{2}\left(1+(2F(t)-1)^{2,5}\right)}-e^{-\theta\left(1+(2F(t)-1)^{2,5}\right)}}{-e^{-\theta}+2e^{-\frac{\theta}{2}(1+(2F(t)-1)^{2,5})}-e^{-\theta(1+(2F(t)-1)^{2,5})}}\right)f(t)$$
$$+3(2G(t)-1)^{2}\left(1-\frac{e^{-\frac{\theta}{2}\left(1+(2G(t)-1)^{3}\right)}-e^{-\theta\left(1+(2G(t)-1)^{3}\right)}}{-e^{-\theta}+2e^{-\frac{\theta}{2}(1+(2G(t)-1)^{3})}-e^{-\theta(1+(2G(t)-1)^{3})}}\right)g(t). \quad (4.38)$$

A Figura 4.6 mostra nos gráficos (a), (b) e (c), a maior ponderação da região das caudas das duas distribuições componentes. A participação que cada cauda das componentes tem no modelo resultante tende à maior ponderação das caudas à esquerda das componentes quando a dependência é negativa e também tende a se equilibrar conforme aumenta a dependência da cópula até τ de Kendall próximo a 1. Quando a dependência é positiva, esse equilíbrio acompanha do padrão de assimetria da distribuição componente, mantendo pesos maiores para a cauda mais pesada.

A Figura 4.7 mostra nos gráficos (a), (b) e (c), as funções densidade, taxa de falha e sobrevivência do modelo resultante dos níveis de dependência da cópula. O modelo mistura valores das regiões das caudas das duas distribuições Weibull gerando funções densidade e de taxa de falha multimodais com a tendência de ponderar mais as caudas à esquerda e menos as caudas à direita das componentes para dependência negativa. Uma situação oposta, para maior ponderação das caudas à direita quando a dependência é negativa, pode ser construída tomando-se as funções de seleção $r(u,v)=[1-(2u-1)^{1/k_1}]/2$ e $s(u,v)=[1-(2v-1)^{1/k_2}]/2$. ■

Exemplo 13 *Considere a especificação do modelo MGC-3CS-CC, do Exemplo 6 com componentes Weibull e cópula de Frank. Nesse caso, temos que a função de distribuição (4.2) é dada por*

$$P(T\leq t)=F(t)+\frac{1}{\theta}\log\left\{1-\frac{(1-e^{-\theta F(t)})(1-e^{-\theta q})}{(1-e^{-\theta})}\right\}-\frac{1}{\theta}\log\left\{1-\frac{(1-e^{-\theta F(t)})(1-e^{-\theta p})}{(1-e^{-\theta})}\right\}$$
$$-\frac{1}{\theta}\log\left\{1-\frac{(1-e^{-\theta G(t)})(1-e^{-\theta q})}{(1-e^{-\theta})}\right\}+\frac{1}{\theta}\log\left\{1-\frac{(1-e^{-\theta G(t)})(1-e^{-\theta p})}{(1-e^{-\theta})}\right\} \quad (4.39)$$

e a função densidade, por

$$\tilde{f}(t)=\left\{1-\frac{e^{-\theta F(t)}(1-e^{-\theta q})}{e^{-\theta F(t)}+e^{-\theta q}-e^{-\theta}-e^{\theta[F(t)+q]}}+\frac{e^{-\theta F(t)}(1-e^{-\theta p})}{e^{-\theta F(t)}+e^{-\theta p}-e^{-\theta}-e^{\theta[F(t)+p]}}\right\}f(t)$$
$$+\left\{\frac{e^{-\theta G(t)}(1-e^{-\theta q})}{e^{-\theta G(t)}+e^{-\theta q}-e^{-\theta}-e^{\theta[G(t)+q]}}-\frac{e^{-\theta G(t)}(1-e^{-\theta p})}{e^{-\theta G(t)}+e^{-\theta p}-e^{-\theta}-e^{\theta[G(t)+p]}}\right\}g(t). \quad (4.40)$$

As Figuras 4.6 e 4.7 ilustram a generalização do modelo de misturas tradicional pelo modelo (4.39) com as distribuições F e G Weibull dadas, cópula de Frank e constantes $p=0,1$ e $q=0,7$. Para a situação de dependência alta positiva a mistura toma aproximadamente as regiões de F dadas pelos quantis correspondentes a A, isto é, até o quantil de probabilidade $0,1$ e após o quantil de probabilidade $0,7$, e de G entre os quantis de probabilidade $0,1$ e $0,7$. Essa disposição é invertida para a situação de dependência alta e negativa, em que participam da mistura as regiões de F até aproximadamente o quantil de probabilidade $0,3$ e após o quantil de probabilidade $0,9$ e de G a região entre os quantis de $0,3$ e $0,9$, aproximadamente. Na situação de independência o modelo equivale ao de mistura tradicional com pesos constantes $1+p-q=0,4$ para F e $q-p=0,6$ para G. ■

4.3 Mistura com Mais de Duas Distribuições

Nesta seção são apresentadas algumas formas de construir modelos de misturas de k distribuições utilizando o modelo MGC. Essas misturas podem ser obtidas por meio da definição dos conjuntos de misturas em um hipercubo unitário, $[0,1]^k$, sobre o qual é definida uma cópula k-variada, ou por meio da definição dos conjuntos de misturas dados pela partição de $[0,1]^2$ em k regiões, ou ainda, pela aplicação de misturas sucessivas da mistura de duas distribuições definida em (4.1).

O exemplo a seguir ilustra uma aplicação do modelo de misturas a três distribuições utilizando uma cópula simétrica definida no cubo unitário. Consideramos uma cópula trivariada simétrica como a cópula que tem argumentos intercambiáveis, isto é, $C(u,v,w) = C(u,w,v) = C(w,v,u) = C(v,u,w)$. Devido à simetria dos conjuntos da mistura em relação aos planos que seccionam o cubo unitário em $u = v$, $v = w$ e $u = w$, esse modelo é incluso na familia de modelos MGC-CR. A utilização de cópulas simétricas nessa abordagem simplifica a função de distribuição do modelo, tornando direta a sua construção, como mostrado no exemplo a seguir.

Exemplo 14 *Considere U, V e W três variáveis aleatórias uniformes em $[0,1]$ tais que*

$$(U,V,W) \sim C, \tag{4.41}$$

em que C é uma função de cópula trivariada simétrica. Para F, G e H funções de distribuição contínuas dadas, definimos a variável T por

$$T = F^{-1}(U)I_{[U\leq V,U\leq W]} + G^{-1}(V)I_{[V<U,V\leq W]} + H^{-1}(W)I_{[W<U,W<V]}, \tag{4.42}$$

em que os conjuntos da mistura são as três pirâmides oblíquas no cubo unitário $[0,1]^3$ de vértices em $(1,1,1)$ e bases contidas nos planos $u = 0$, $v = 0$ e $w = 0$. Os elementos (u,v,w) do cubo tais que u=v e u=w estão contidos na primeira pirâmide e os elementos tais que v=w estão contidos na segunda pirâmide. Temos então que o quantil de probabilidade u de F, v de G, ou w de H, é tomado para T, conforme a ocorrência da realização (u,v,w) de $(U,V,W) \sim C$ nos conjuntos da mistura.

A função de distribuição de T é dada por

$$\begin{aligned}
P[T \leq t] &= 1 - P[F^{-1}(U)I_{[U\leq V,U\leq W]} + G^{-1}(V)I_{[V<U,V\leq W]} + H^{-1}(W)I_{[W<U,W<V]} > t] \\
&= 1 - P[F^{-1}(U) > t, U \leq V, U \leq W] + P[G^{-1}(V) > t, V < U, V \leq W] \\
&\quad + P[H^{-1}(W) > t, W < U, W < V] \\
&= 1 - P[U > F(t), U \leq V, U \leq W] + P[V > G(t), V < U, V \leq W] \\
&\quad + P[W > H(t), W < U, W < V] \\
&= F(t) - \delta_C(F(t)) + \gamma_C(F(t))/3 + G(t) - \delta_C(G(t)) + \gamma_C(G(t))/3 \\
&\quad H(t) - \delta_C(H(t)) + \gamma_C(H(t))/3,
\end{aligned} \tag{4.43}$$

em que $\gamma_C(x) = C(x,x,x)$ é a diagonal da cópula trivariada. A função densidade é dada por

$$\begin{aligned}
P[T \leq t] &= [1 - \delta'_C(F(t)) + \gamma'_C(F(t))/3]f(t) + [1 - \delta'_C(G(t)) + \gamma'_C(G(t))/3]g(t) \\
&\quad + [1 - \delta'_C(H(t)) + \gamma'_C(H(t))/3]h(t),
\end{aligned} \tag{4.44}$$

em que $\gamma'_C(x) = \frac{d}{dx}C(x,x,x)$. Nesse caso, pela simetria dos conjuntos no cubo unitário, temos $P[U = Z] = P[V = Z] = P[W = Z] = 1/3$ é a probabilidade da tomada de percentis F, G e H. ∎

Uma outra forma de definir uma mistura de k distribuições utilizando cópulas é definir os conjuntos da mistura como uma partição de k regiões de $[0,1]^2$. Para a mistura de k distribuições componentes

podemos definir

$$T = \sum_{j=1}^{k} F_j^{-1}(r_j(U,V))I_{(U,V)\in A_j} \tag{4.45}$$

em que F_j são as distribuições componentes, A_j são os conjuntos da mistura e $r_j(u,v)$ são as funções de seleção, para $j = 1, ..., k$. Os conjuntos da mistura podem ser definidos de forma que, para $j = 1, ..., k$, tenhamos $r_j(u,v) = v$ e $A_j = [u_{j-1}, u_j)$, que utilizam as constantes $u_0 = 0$, $u_k = 1$ e as constantes u_j, $j = 1, .., k-1$ que são $k-1$ parâmetros da mistura nessa família. Essa família é considerada uma família de modelos de conjuntos simples devido ao tipo de partição realizada em $[0,1]^2$. O exemplo a seguir considera o caso de $k = 3$.

Exemplo 15 *Considere as funções de distribuição F_1, F_2 e F_3, as funções de seleção $r_1(u,v)$, $r_2(u,v)$ e $r_3(u,v)$ e também os conjuntos de mistura A_1, A_2 e A_3 dados por uma partição de $[0,1]^2$. Definimos a variável*

$$T = F_1^{-1}(r_1(U,V))I_{(U,V)\in A_1} + F_2^{-1}(r_2(U,V))I_{(U,V)\in A_2} + F_3^{-1}(r_3(U,V))I_{(U,V)\in A_3} \tag{4.46}$$

do que segue a função de distribuição

$$\begin{aligned} P[T \le t] &= P[r_1(U,V) \le F_1(t), (U,V) \in A_1] + P[r_2(U,V) \le F_2(t), (U,V) \in A_2] \\ &\quad + P[r_3(U,V) \le F_3(t), (U,V) \in A_3] \\ &= \int\int_{[r_1(U,V)\le F_1(t),(U,V)\in A_1]} dC(u,v) + \int\int_{[r_2(U,V)\le F_2(t),(U,V)\in A_2]} dC(u,v) \\ &\quad + \int\int_{[r_3(U,V)\le F_3(t),(U,V)\in A_3]} dC(u,v). \end{aligned} \tag{4.47}$$

O modelo apresentado é uma generalização do modelo de MGC-3CS-CC analisado em (4.30), em que $r_1 = r_3 = r = v$, $r_2 = s = v$, $A_1 = [u \le p]$, $A_2 = [p \le u < q]$ $A_3 = [u \ge q]$, F_1 e F_3 não necessariamente são a mesma função de distribuição F. Nesse caso, temos $P[A_1]$, $P[A_2]$ e $P[A_3]$ as probabilidades de se tomar para a mistura valores de F_1, F_2 e F_3, respectivamente. O Exemplo 13 do modelo MGC-3CS-CC serve de ilustração para mostrar que o tratamento dessa abordagem de mistura de k distribuições é direto, de forma equivalente à mistura feita para duas distribuições. ■

Uma mistura de k distribuições também pode ser construída a partir de misturas sucessivas de duas distribuições, conforme o exemplo ilustra a seguir para $k = 3$.

Exemplo 16 *Para um modelo de misturas definido como em (4.1), considere o vetor aleatório $(U_1, V_1) \sim C$, as funções de seleção em $(0,1)$ $r_1(u,v)$ e $s_1(u,v)$, as funções de distribuição F e G e os conjuntos da mistura A_1 e A_1^c dados. Definda a variável*

$$T = F^{-1}(r_1(U_1,V_1))I_{(U_1,V_1)\in A_1} + G^{-1}(s_1(U_1,V_1))I_{(U_1,V_1)\in A_1^c}, \tag{4.48}$$

segue que a função de distribuição de T é dada por

$$\begin{aligned} \tilde{F}(t) &= P[T \le t] \\ &= P[r_1(U_1,V_1) \le F(t), (U_1,V_1) \in A_1] + P[s_1(U_1,V_1) \le G(t), (U_1,V_1) \in A_1^c] \\ &= \int\int_{[r_1(U_1,V_1)\le F(t),(U_1,V_1)\in A_1]} dC(u,v) + \int\int_{[s_1(U_1,V_1)\le G(t),(U_1,V_1)\in A_1^c]} dC(u,v). \end{aligned} \tag{4.49}$$

Considere também o vetor aleatório $(U_2, V_2) \sim C$*, as funções de seleção em* $(0,1)$ $r_2(u,v)$ *e* $s_2(u,v)$*, as funções de distribuição* $\tilde{F}$ *e* H *e os conjuntos da mistura* A_2 *e* A_2^c *dados. Definimos a variável*

$$W = \tilde{F}^{-1}(r_2(U_2,V_2))I_{(U_2,V_2)\in A_2} + H^{-1}(s_2(U_2,V_2))I_{(U_2,V_2)\in A_2^c} \tag{4.50}$$

segue que a função de distribuição de W *é dada por*

$$\begin{aligned} P[W \leq t] &= P[r_2(U_2,V_2) \leq \tilde{F}(t), (U_2,V_2) \in A_2] + P[s_2(U_2,V_2) \leq H(t), (U_2,V_2) \in A_2^c] \\ &= \int\int_{[r_2(U_2,V_2)\leq \tilde{F}(t), (U_2,V_2)\in A_2]} dC(u,v) + \int\int_{[s_2(U_2,V_2)\leq H(t), (U_2,V_2)\in A_2^c]} dC(u,v) \end{aligned} \tag{4.51}$$

■

4.4 Aplicações

Nesta seção algumas famílias do modelo de mistura generalizado apresentadas são ajustadas a conjutos de dados simulados e empíricos da literatura. Os parâmetros dos modelos propostos foram estimados pelo método de máxima verossimilhança. Considerando uma amostra $x_i = (\delta_i, t_i), i = 1, \cdots, n$, com censuras aleatórias à direita para as quais δ_i é variável indicadora de falha e t_i é o mínimo entre valor observado e o tempo de censura, segue de (4.2) e (4.3) que a função de verossimilhança é dada por

$$L(\Upsilon; x) = \prod_{i=1}^{n} \tilde{f}(t_i;\Upsilon)^{\delta_i} \tilde{S}(t_i;\Upsilon)^{1-\delta_i}, \tag{4.52}$$

em que Υ denota os parâmetros do modelo de mistura. Para o ajuste dos modelos a função de log-verossimilhança foi otimizada. Foram utilizados algoritmos escritos em R e a função *optim* com o método combinado de dois passos sendo o primeiro de 'Nelder-Mead', seguido por 'BFGS' executado a partir do resultado do primeiro passo. No procedimento de otimização foram também analisados os problemas de máximos locais e identificação dos modelos. Para isso, no ajuste de cada modelo proposto foram utilizados de 300 a 600 valores iniciais tomados de um hiper-retângulo do espaço paramétrico do modelo. Os resultados da otimização dos modelos propostos em cada amostra estudada foram analisados simultaneamente para verificar a qualidade geral dos ajustes e a presença de máximos locais. Nas aplicações realizadas os resultados mostraram a presença de máximos locais e fortes evidências que os modelos analisados são identificáveis. Em cada modelo ajustado os pontos encontrados no espaço paramétrico em que a função de log-verossimilhança obteve valor máximo foram iguais, exceto os casos das famílias de conjuntos simples MGC-CS-O e MGC-3CS-CC com duas componentes de mesma especificação, que permitem duas parametrizações atingirem o máximo, conforme discutido nos Exemplos 4 e 6. Para ilustrar essa análise, a Figura 4.8 mostra o resultado do procedimento de otimização da função de log-verossimilhança para os modelos propostos da família MG-CR-EE e de misturas tradicional ajustados a dados simulados do modelo MGC-CR-EE, com componentes Weibull e cópula de Frank. Os modelos propostos utilizam várias componentes combinadas para essa mesma família de mistura com cópula de Frank ou para o modelo de misturas tradicional, como mostrado no gráfico. Para cada modelo MGC a otimização utilizou 600 valores iniciais e de misturas tradicional, 300 valores iniciais. Cada ponto do gráfico mostra o resultado da otimização da função a partir de um ponto inicial distinto e é exibido com um erro lateral aleatório a fim de que o gráfico informe a densidade de cálculos de otimização da função com mesmo resultado do valor da log-verossimilhança. Para cada modelo proposto, os valores concentrados em determinado valor da função de log-verossimilhança indicam máximos locais ou o máximo global. Os valores não pertencentes a essas regiões de concentração de pontos mostram situações em que houve falha de convergência do algoritmo ou que o algoritmo buscou as estimativas em regiões

do espaço paramétrico em que a estabilidade numérica do cálculo da função de log-verossimilhança é ruim. Por exemplo, se o τ de Kendall é muito próximo a $|1|$, o parâmetro da cópula de Frank θ, que é argumento de funções exponenciais da função de cópula, assume valores altos e prejudica a estabilidade numérica. Nos casos mencionados acima foi observado $\tau > 0,99$ e, nesse caso, temos $\theta > 398,34$. Os valores da função para as soluções não factíveis, i.e. não pertencentes ao espaço paramétrico, não são exibidas no gráfico. Foi também verificado que as soluções em que a função atingiu o valor máximo são o mesmo ponto do espaço paramétrico, o que indica a identificabilidade dos modelos analisados. O máximo da função foi atingido em cerca de 60% dos valores iniciais para o modelo MGC de melhor ajuste, de ambas componentes Weibull, e 76% para o modelo de mistura tradicional de melhor ajuste, também de ambas componentes Weibull. Esse percentual variou de 2,3% (MTrad Lnor-Lnor) a 96,7% (MTrad Llog-Llog) entre os modelos do gráfico.

4.4.1 Dados Simulados

O procedimento de estimação descrito foi aplicado a dados simulados de três modelos MGC entre os apresentados na seção 4.2. De cada modelo foram simuladas observações independentes formando três amostras de tamanhos $n = 100$, 300 e 1.000. Em seguida foram estimados para cada amostra modelos MGC propostos com distribuições componentes log-logística, log-normal, gama e Weibull. O modelo de misturas tradicional também foi estimado com essas componentes para comparação. Utilizamos nessa seção as abreviações Llog, Lnor, Gam e Wei para as distribuições log-logistica, log-normal, gama e Weibull, respectivamente, para fazer referência a uma especificação do modelo de misturas. Por exemplo, em uma família de modelo de misturas pode-se especificar o uso da função de cópula de Frank e distribuições componentes Weibull por Frank-Wei-Wei. Como os casos simulados utilizam distribuições de variáveis aleatórias positivas, os resultados foram analisados segundo os procedimentos de análise de sobrevivência, envolvendo a taxa de falha e a função de sobrevivência. A cada amostra foram aplicadas censuras aleatórias na proporção de 10%. O ajuste dos modelos foi comparado segundo o critério de Akaike (Akaike, 1973) e o critério de informação Bayesiano (Schwarz, 1978). O critério de Akaike foi calculado como $AIC = -2L(\hat{\Upsilon}) + 2k'$, para k' o número de parâmetros e $L(\hat{\Upsilon})$ a função de log-verossimilhança calculada na estimativa de máxima verossimilhança $\hat{\Upsilon}$ e o critério de informação Bayesiano, como $BIC = -2L(\hat{\Upsilon}) + \log(n)k'$, em que n é o tamanho da amostra. Nos conjuntos de dados simulados dessa seção os critérios AIC e BIC levaram às mesmas decisões de comparação dos modelos. Dessa forma, para cada modelo estimado, é apenas mostrado o critério AIC.

Caso 1: No primeiro caso analisado os dados foram simulados do modelo MGC-3CS-CC, do Exemplo 13. Para esse modelo, definimos o conjunto de parâmetros $\Upsilon = (p, k, \theta, \mu_1, \beta_1, \mu_2, \beta_2)$ em que os parâmetros dos conjuntos de mistura são $p = 0,1$ e $k = 0,6$; o parâmetro da cópula de Frank é $\theta = 18,19$, o que equivale ao τ de Kendall igual a $0,80$; e os parâmetros das distribuições componenetes Weibull são $(\mu_1, \beta_1) = (2,0; 3,0)$ e $(\mu_2, \beta_2) = (7,0; 6,0)$. Na aplicação, trabalhamos com a reparametrização dada pela diferença entre q e p, $k = q - p$, dos parâmetros apresentados na apresentação do modelo na Seção 4.1. A Tabela 4.1 apresenta, segundo o critério AIC, as estimativas dos modelos MGC-3CS-CC de melhor ajuste aos dados das amostras de $n = 100$, 300 e 1.000 e o modelo de misturas tradicional proposto de melhor ajuste. Os intervalos de confiança para os parâmetros são exibidos em parênteses e foram calculados numericamente pela informação de Fisher. O intervalo para o τ de Kendall foi calculado como função dos limites do intervalo para o parâmetro de dependência. Os modelos de mistura generalizados obtiveram melhor ajuste que os modelos de mistura tradicional em geral. Também, nas três amostras o modelo proposto com a especificação correta foi o de melhor critério AIC. A Figura 4.9 apresenta os gráficos das funções densidade, taxa de falha e sobrevivência dos modelos estimados para a amostra de tamanho $n = 300$ em comparação com o modelo real e os modelos de misturas generalizados exibem muito bom ajuste. A Tabela 4.1 também mostra a melhoria da precisão das estimativas

dos parâmetros conforme aumenta o tamanho da amostra. Para o modelo ajustado com a especificação Frank-Gam-Wei foram também obtidos com mesmo valor de log-verossimilhança pontos de máximo em $\Upsilon = (0,33; 0,61; -0,81; 6,62; 0,28; 7,11; 6,93)$ para a amostra de tamanho $n = 300$, conforme o problema de identificação descrito para os modelos da família MGC-3CS-CC ao final do Exemplo 6.

Tabela 4.1: *Informações do ajuste dos modelos para os dados simulados no Caso 1 de modelo MGC-3CS-CC e tamanhos de amostra n=100, 300 e 1.000. São exbidos os modelos de melhor critério AIC e o modelo de misturas tradicional de melhor ajuste.*

n=100									
Modelo	AIC	Mistura		τ	θ	Componente 1		Componente 2	
MGC-3CS-CC-Frank-Wei-Wei	299,30	0,06 (0,02;0,11)	0,56 (0,46;0,65)	0,76 (0,59;0,84)	15,03 (7,55;22,51)	2,07 (1,93;2,21)	3,32 (2,68;3,96)	7,05 (6,74;7,36)	8,12 (5,23;11,00)
MGC-3CS-CC-Frank-Gam-Gam	306,07	0,06 (0,01;0,10)	0,55 (0,46;0,65)	0,81 (0,66;0,87)	18,82 (9,62;28,03)	7,17 (4,55;9,80)	0,26 (0,16;0,36)	36,26 (13,39;59,13)	0,19 (0,07;0,31)
MGC-3CS-CC-Frank-Gam-Wei	306,25	0,06 (0,02;0,11)	0,57 (0,47;0,67)	0,79 (0,65;0,85)	17,07 (9,45;24,69)	7,30 (4,69;9,92)	0,25 (0,16;0,35)	7,04 (6,74;7,35)	7,97 (5,09;10,85)
MTrad-Wei-Wei	321,56	0,46 (0,36;0,56)				2,45 (2,27;2,63)	4,21 (3,14;5,29)	6,55 (6,36;6,74)	10,75 (8,40;13,09)
n=300									
Modelo	AIC	Mistura		τ	θ	Componente 1		Componente 2	
MGC-3CS-CC-Frank-Wei-Wei	940,61	0,06 (0,03;0,09)	0,61 (0,55;0,66)	0,79 (0,72;0,83)	16,80 (12,20;21,41)	2,07 (1,97;2,16)	3,00 (2,63;3,36)	7,10 (6,92;7,28)	6,89 (5,67;8,11)
MGC-3CS-CC-Frank-Gam-Wei	944,51	0,06 (0,03;0,09)	0,61 (0,55;0,67)	0,81 (0,73;0,86)	19,39 (12,89;25,89)	6,62 (5,16;8,08)	0,28 (0,22;0,34)	7,11 (6,93;7,29)	6,93 (5,74;8,13)
MGC-3CS-CC-Frank-Gam-Gam	946,05	0,06 (0,03;0,08)	0,60 (0,54;0,65)	0,82 (0,76;0,86)	20,69 (14,66;26,72)	6,49 (5,06;7,92)	0,29 (0,22;0,36)	29,84 (20,61;39,07)	0,23 (0,16;0,30)
MTrad-Wei-Wei	997,59	0,60 (0,54;0,66)				2,57 (2,42;2,73)	3,44 (2,84;4,04)	6,56 (6,45;6,67)	10,27 (8,89;11,64)
n=1.000									
Modelo	AIC	Mistura		τ	θ	Componente 1		Componente 2	
MGC-3CS-CC-Frank-Wei-Wei	3150,46	0,10 (0,08;0,12)	0,60 (0,57;0,63)	0,79 (0,76;0,81)	17,02 (14,52;19,52)	2,01 (1,96;2,06)	3,07 (2,89;3,26)	7,00 (6,89;7,10)	6,20 (5,59;6,80)
MGC-3CS-CC-Frank-Gam-Wei	3171,92	0,10 (0,08;0,12)	0,60 (0,57;0,63)	0,81 (0,77;0,84)	19,37 (15,92;22,81)	6,61 (5,91;7,32)	0,27 (0,24;0,30)	7,01 (6,91;7,11)	6,18 (5,59;6,77)
MGC-3CS-CC-Frank-Gam-Gam	3180,55	0,09 (0,08;0,11)	0,59 (0,56;0,62)	0,82 (0,79;0,84)	20,04 (17,14;22,94)	6,53 (5,83;7,22)	0,28 (0,25;0,31)	25,12 (20,67;29,56)	0,27 (0,22;0,32)
MTrad-Wei-Wei	3379,11	0,59 (0,56;0,63)				2,40 (2,31;2,50)	2,80 (2,53;3,08)	6,54 (6,48;6,60)	9,95 (9,22;10,67)

Caso 2: No segundo caso os dados foram simulados do modelo MGC-CR-CAE, do Exemplo 12. Para esse modelo definimos os parâmetros por $\Upsilon = (k_1, k_2, \theta, \mu_1, \beta_1, \mu_2, \beta_2)$ em que os parâmetros das funções de seleção são $k_1 = 2,5$ e $k_2 = 3,0$; o parâmetro da cópula de Frank é $\theta = 18,19$, para τ de Kendall igual a $0,80$; e os parâmetros das distribuições componentes Weibull são $(\mu_1, \beta_1) = (2,0; 3,0)$ e $(\mu_2, \beta_2) = (7,0; 6,0)$. A Tabela 4.2 apresenta as estimativas dos modelos MGC-CR-CAE de melhor ajuste, pelo critério AIC, ajustados às amostras de $n = 100$, 300 e 1.000 e também o modelo de misturas tradicional de melhor ajuste em cada caso de tamanho de amostra para comparação. Os parâmetros dos modelos que aparecem nos três casos mostram a melhor precisão das estimativas conforme é aumentado o tamanho da amostra. A Figura 4.10 mostra as funções densidade, taxa de falha e sobrevivência da Tabela 4.2 e amostra $n = 300$. Os modelos MGC exibem bom ajuste às quatro modas dos dados pela função densidade e também representam bem a função taxa de falha nessas condições.

Caso 3: No terceiro caso analisado as amostras foram geradas do modelo MGC-CR-EE, do Exemplo 7. Para esse modelo definimos os parâmetros por $\Upsilon = (\theta, \mu_1, \beta_1, \mu_2, \beta_2)$ em que o parâmetro da cópula de Frank é $\theta = -18,19$, o que equivale ao índice de concordância de τ de Kendall igual a $-0,80$; e os

Tabela 4.2: *Informações do ajuste dos modelos para os dados simulados no Caso 2 de modelo MGC-CR-CAE e tamanhos de amostra n=100, 300 e 1.000. São exibidos os modelos de melhor critério AIC e o modelo de misturas tradicional de melhor ajuste.*

n=100									
Modelo	AIC	Mistura		τ	θ	Componente 1		Componente 2	
MGC-CR-CAE-	388,00	2,11	3,80	0,88	30,66	1,96	2,64	6,96	6,45
Frank-Wei-Wei		(1,31;2,91)	(1,90;5,69)	(-0,95;0,97)	(-75,75;137,07)	(1,84;2,09)	(2,08;3,20)	(6,75;7,16)	(4,89;8,02)
MGC-CR-CAE-	389,04	2,20	54,79	0,68	10,40	15,61	0,25	4,85	3,20
Frank-Gam-Wei		(1,36;3,04)	(1,46;108,12)	(-0,47;0,86)	(-5,20;26,00)	(9,45;21,76)	(0,15;0,35)	(4,62;5,09)	(2,64;3,75)
MGC-CR-CAE-	389,66	2,19	3,64	0,84	23,29	1,66	3,69	6,97	6,33
Frank-Llog-Wei		(1,22;3,16)	(1,73;5,55)	(-0,83;0,94)	(-22,20;68,77)	(1,55;1,78)	(2,63;4,76)	(6,75;7,18)	(4,79;7,87)
MTrad	402,81	0,86				3,44	1,63	8,30	23,07
Wei-Wei		(0,77;0,95)				(2,87;4,00)	(1,32;1,94)	(8,02;8,58)	(8,68;37,46)
n=300									
Modelo	AIC	Mistura		τ	θ	Componente 1		Componente 2	
MGC-CR-CAE-	1097,56	2,68	3,25	0,96	96,87	1,97	3,11	6,91	6,24
Frank-Wei-Wei		(2,06;3,30)	(2,46;4,04)	(-0,95;0,99)	(-77,08;270,82)	(1,91;2,03)	(2,76;3,45)	(6,80;7,01)	(5,52;6,96)
MGC-CR-CAE-	1107,37	2,66	69,12	0,96	100,55	15,57	0,23	4,71	3,25
Frank-Gam-Wei		(2,03;3,30)	(37,64;100,61)	(-0,94;0,99)	(-69,49;270,60)	(11,91;19,23)	(0,18;0,29)	(4,59;4,83)	(2,98;3,52)
MGC-CR-CAE-	1107,89	2,65	56,79	0,82	20,29	1,27	0,25	4,74	3,16
Frank-Lnor-Wei		(2,00;3,30)	(26,80;86,78)	(-0,79;0,93)	(-16,86;57,44)	(1,25;1,30)	(0,22;0,29)	(4,60;4,88)	(2,88;3,45)
MTrad	1166,64	0,25				2,07	0,06	3,17	1,69
Lnor-Wei		(0,19;0,31)				(2,05;2,09)	(0,05;0,08)	(2,86;3,48)	(1,49;1,90)
n=1.000									
Modelo	AIC	Mistura		τ	θ	Componente 1		Componente 2	
MGC-CR-CAE-	3671,47	2,98	2,48	0,84	23,07	1,98	2,92	7,02	5,88
Frank-Wei-Wei		(2,55;3,40)	(2,13;2,84)	(0,30;0,91)	(2,88;43,27)	(1,95;2,02)	(2,69;3,15)	(6,95;7,09)	(5,43;6,33)
MGC-CR-CAE-	3722,36	3,20	2,21	0,84	23,94	31,98	0,20	2,04	2,69
Frank-Gam-Wei		(2,64;3,77)	(1,95;2,48)	(0,59;0,90)	(7,61;40,27)	(26,09;37,87)	(0,17;0,24)	(2,01;2,07)	(2,50;2,87)
MGC-CR-CAE-	3741,53	5,06	45,00	0,74	13,60	4,01	0,65	19,10	0,26
Frank-Gam-Gam		(4,10;6,02)	(29,69;60,31)	(0,50;0,83)	(5,73;21,47)	(3,57;4,46)	(0,58;0,73)	(17,16;21,04)	(0,23;0,29)
MTrad	3883,86	0,18				2,09	0,06	3,29	1,59
Lnor-Wei		(0,15;0,21)				(2,08;2,10)	(0,05;0,07)	(3,11;3,48)	(1,49;1,68)

parâmetros das distribuições componentes Weibull são $(\mu_1, \beta_1) = (2,0; 3,0)$ e $(\mu_2, \beta_2) = (7,0; 6,0)$. A Tabela 4.3 apresenta as estimativas para os modelos MGC-CR-EE de melhor ajuste e o melhor modelo de misturas tradicional para as amostras de $n = 100$, 300 e 1.000, exceto o modelo MGC-CR-EE com especificação Frank-Wei-Wei de nono melhor ajuste, que foi incluso nos modelos da amostra de tamanho $n = 300$ para comparação com os ajustes dessa especificação analisados para as outras amostras. Para a amostra de tamanho $n = 100$ foi necessário restringir o parâmetro da cópula de tal forma que $\tau \leq 0,90$, pois as estimativas tenderam à fronteira $\tau = 1$. Também nesse caso os modelos MGC obtiveram melhor ajuste que os modelos de mistura tradicional. A Figura 4.11 mostra as funções densidade, taxa de falha e sobrevivência para os modelos estimados para a amostra de tamanho $n = 300$. O gráfico mostra as diferenças entre os ajustes para caso de o maior número de amostra, em que a diferença foi mais nítida.

4.4.2 Dados Empíricos

Nesta seção alguns dados clássicos da literatura de modelos de misturas são analisados segundo as famílias de modelos de misturas generalizados propostos esse trabalho. O primeiro caso é o conjunto de dados sobre propriedades químicas da água de lagos nos EUA; o segundo, sobre a presença de enzimas na urina de pacientes em um teste de estudo sobre o câncer; e o terceiro, sobre tempos de erupção do gêiser The Old Faithful. A seguir são ajustados a esses conjuntos de dados os modelos de misturas generalizados com especificações dadas pela cópula de Frank e combinações duas a duas das componentes: gama, log-logística, log-normal e Weibull. Além dessas componentes também é considerada a

Tabela 4.3: *Informações do ajuste dos modelos para os dados no Caso 3 de modelo MGC-CR-EE e tamanhos de amostra n = 100, 300 e 1.000. São exibidos os modelos de melhor critério AIC e o modelo de misturas tradicional de melhor ajuste, exceto o modelo com especificação Frank-Wei-Wei de nono melhor ajuste, que foi incluso na amostra para comparação com os ajustes dessa especificação analisados para as outras amostras.*

$n = 100$								
Modelo	AIC	Mistura	τ	θ	Componente 1		Componente 2	
MGC-CR-EE-Frank-Wei-Wei	280,97		-0,90 (-0,96;0,81)	-38,15 (-95,54;19,24)	2,17 (2,01;2,33)	2,49 (1,86;3,12)	7,00 (6,74;7,26)	4,84 (3,24;6,45)
MGC-CR-EE-Frank-Llog-Wei	281,87		-0,90 (-0,95;0,46)	-38,28 (-81,56;4,99)	1,88 (1,78;1,98)	2,91 (2,23;3,59)	7,00 (6,74;7,26)	4,85 (3,25;6,44)
MGC-CR-EE-Frank-Gam-Wei	282,66		-0,90 (-0,95;0,41)	-38,28 (-80,84;4,27)	3,61 (2,32;4,89)	0,57 (0,34;0,81)	7,00 (6,74;7,26)	4,84 (3,24;6,44)
MTrad-Wei-Wei	294,93	0,59 (0,49;0,69)			1,44 (1,31;1,57)	3,25 (2,38;4,13)	5,70 (5,43;5,97)	7,58 (5,17;9,98)
$n = 300$								
Modelo	AIC	Mistura	τ	θ	Componente 1		Componente 2	
MGC-CR-EE-Frank-Gam-Lnor	825,42		-0,84 (-0,89;-0,66)	-23,01 (-36,27;-9,75)	16,19 (11,81;20,57)	0,42 (0,30;0,53)	0,62 (0,58;0,65)	0,52 (0,45;0,58)
MGC-CR-EE-Frank-Llog-Lnor	825,56		-0,87 (-0,93;-0,30)	-29,93 (-56,89;-2,97)	6,62 (6,50;6,73)	6,36 (5,33;7,39)	0,62 (0,58;0,65)	0,52 (0,45;0,58)
MGC-CR-EE-Frank-Wei-Wei	829,10		-0,78 (-0,88;-0,10)	-16,44 (-32,00;-0,87)	2,04 (1,94;2,14)	3,05 (2,55;3,55)	6,98 (6,78;7,19)	5,69 (4,61;6,77)
MTrad-Wei-Wei	841,54	0,52 (0,46;0,57)			5,78 (5,65;5,90)	8,00 (6,87;9,13)	1,43 (1,37;1,48)	4,18 (3,63;4,73)
$n = 1.000$								
Modelo	AIC	Mistura	τ	θ	Componente 1		Componente 2	
MGC-CR-EE-Frank-Wei-Wei	2595,43		-0,75 (-0,80;-0,67)	-13,98 (-17,73;-10,24)	6,98 (6,89;7,08)	5,95 (5,38;6,52)	1,94 (1,89;1,99)	3,25 (2,96;3,53)
MGC-CR-EE-Frank-Llog-Wei	2597,72		-0,79 (-0,83;-0,72)	-17,48 (-22,32;-12,64)	1,75 (1,72;1,79)	3,72 (3,42;4,01)	7,02 (6,93;7,11)	5,81 (5,26;6,36)
MGC-CR-EE-Frank-Llog-Llog	2598,41		-0,82 (-0,86;-0,76)	-20,59 (-26,53;-14,65)	6,62 (6,56;6,68)	6,71 (6,13;7,29)	1,76 (1,73;1,80)	3,66 (3,37;3,95)
MTrad-Wei-Wei	2669,84	0,48 (0,45;0,52)			5,83 (5,75;5,90)	8,04 (7,40;8,68)	1,39 (1,37;1,42)	4,47 (4,14;4,79)

especificação de duas distribuições normais, para comparação com outros resultados da literatura que serão discutidos. As mesmas especificações de distribuições componentes também são estimadas com o modelo de misturas tradicional para comparação.

Acidez da água em lagos

O conjunto de dados sobre a acidez de 155 lagos no estado de Wisconsin, EUA, é amplamente analisado na literatura de modelos de misturas. Crawford *et al.* (1992) e Crawford (1994) propõem uma análise Bayesiana do índice de acidez por meio de duas distribuições componentes log-normais em que os indivíduos têm pesos diferentes com base em suas covariáveis. Richardson e Green (1997) produziu uma abordagem por MCMC com saltos reversíveis para modelos de mistura de distribuições normais univariadas com um número desconhecido de componentes e identificou para os dados de acidez em lagos entre 2 e 6 o número mais razoável de componentes de uma mistura de distribuições normais. Em Mclachlan e Peel (2000) a inferência para o número de componentes por meio de bootstrap da função de log-verossimilhança indica duas componentes.

A análise Bayesiana de misturas de distribuições de Crawford *et al.* (1992) pelo método de Laplace tem o objetivo de obter estimativas dos parâmetros do modelo de misturas e probabilidades de classificação dos lagos nas componentes. Os dados da medida de acidez dos lagos são modelados por uma mistura de duas populações lognormais, usando outras características dos lagos como variáveis

explicativas para classificação dos lagos em sub-populações. Para cada lago é estimada a probabilidade de classificação em uma das componentes dadas por uma função logística das características dos lagos. Essas probabilidades são utilizadas como pesos individuais para cada lago para a proporção da primeira componente e o modelo de misturas ajustado mostra uma boa representação da distribuição empírica observada.

Os dados foram analisados com os modelos de misturas generalizados da família MGC-CS-O (Exemplo 4) e pelos modelos de misturas tradicional com as combinações de distribuições componentes citadas. Conforme as especificações de distribuições componentes mencionadas com cópula de Frank, resultam 11 modelos de mistura generalizados e 11 de misturas tradicional. A Tabela 4.4 mostra o ajuste dos três modelos de misturas que apresentaram melhor ajuste pelo critério AIC, que também foram os de melhor BIC. Também, para comparação, é mostrado o modelo de misturas tradicional de melhor ajuste pelo critério AIC, de componentes log-logística e Weibull. Esse modelo obteve nono melhor AIC e quinto melhor BIC, considerando todos os modelos. Segundo as estimativas do parâmetro p, os modelos de mistura generalizados atribuíram probabilidade de aproximadamente 50% para cada uma de suas parcelas enquanto que o modelo de misturas tradicional atribuiu 58% e 42% para as suas componentes. A dependência negativa da cópula nos três modelos de misturas generalizados mostrados na tabela indica que os modelos de mistura tomam valores das regiões à direita da primeira componente e à esquerda da segunda componente. Também a componente Weibull nessas três especificações foi estimada com parâmetros relativamente próximos. As funções densidade dos modelos selecionados são mostradas na Figura 4.12, em que os modelos de misturas generalizados exibem flexibilidade para representar a assimetria contida nos dados.

Tabela 4.4: *Critérios de ajuste AIC e BIC e parâmetros dos modelos de mistura da família MGC-CS-O e mistura tradicional estimados para o conjunto de dados de acidez da água em lagos. São mostrados os modelos de melhor ajuste pelo critério AIC e também o modelo de misturas tradicional de melhor ajuste para comparação.*

Modelo	AIC	BIC	Mistura	τ	θ	Componente 1		Componente 2	
MGC-CS-O-Frank-Wei-Wei	364,82	383,08	0,51 (0,40;0,62)	-0,85 (-0,91;-0,49)	-24,82 (-44,01;-5,62)	4,10 (3,97;4,24)	7,75 (5,93;9,58)	7,36 (7,04;7,69)	6,94 (4,46;9,43)
MGC-CS-O-Frank-Gam-Wei	365,71	383,97	0,53 (0,41;0,64)	-0,81 (-0,89;-0,51)	-19,70 (-33,50;-5,90)	64,27 (36,92;91,62)	0,06 (0,04;0,09)	7,36 (7,03;7,69)	7,49 (4,55;10,43)
MGC-CS-O-Frank-Lnor-Wei	365,88	384,14	0,53 (0,42;0,65)	-0,80 (-0,88;-0,44)	-18,35 (-31,89;-4,81)	1,38 (1,35;1,41)	0,12 (0,09;0,15)	7,35 (7,01;7,70)	7,60 (4,55;10,66)
MTrad-Llog-Wei	370,45	385,66	0,58 (0,48;0,68)			4,30 (4,21;4,39)	21,31 (16,12;26,49)	6,44 (6,27;6,62)	12,84 (8,64;17,04)

A Figura 4.13 mostra a relação dos escores obtidos pelas covariáveis em Crawford *et al.* (1992) com a medida de acidez dos lagos. No gráfico, quando a acidez é baixa a probabilidade de pertencimento à primeira componente é alta e quando a acidez é alta a probabilidade de pertencimento à primeira componente é alta ou baixa. Uma análise semelhante, que relaciona escores baseados nas covariáveis à parametrização do modelo, pode ser realizada para o modelo de misturas generalizados. Por exemplo, na família MGC-CS-O, os parâmetro da cópula e da mistura que definem os graus de assimetria e pesos das componentes podem ser relacionados a covariáveis segundo métodos de análise Bayesiana.

Atividade enzimática em teste para detecção de genótipos associados ao câncer

O segundo conjunto de dados é resultado de um experimento realizado envolvendo 245 indivíduos com o objetivo de analisar a cafeína como droga de prova para detecção de padrões genotípicos associados ao câncer. De cada indivíduo foi observada a razão de quantidades de metabólitos (AFMU/1X) presentes na urina após o indivíduo tomar uma dose de cafeína. Os dados foram analisados inicialmente

por Bechtel *et al.* (1993), em que foi identificada uma mistura de duas distribuições assimétricas utilizando o método de estimação de Maclean *et al.* (1976). Richardson e Green (1997) analisaram esse conjunto de dados segundo um modelo de misturas de distribuições normais estimado pela análise MCMC com saltos reversíveis encontrando evidências estatísticas para a presença de 3 a 6 componentes normais, enquanto que Mclachlan e Peel (2000) indicam uma mistura de 3 componentes normais pelo método de bootstrap da função de log-verossimilhança. Lin *et al.* (2007) analisaram esses dados segundo o modelo de misturas finitas com duas componentes skew normal (SNMIX), que utilizamos em comparação com o ajuste dos modelos de misturas generalizados a seguir.

Os dados foram analisados segundo os modelos de misturas generalizados da família MGC-CS-CED (Exemplo 4) e de misturas tradicional com as combinações de componentes citadas. Conforme as especificações de distribuições componentes mencionadas e cópula de Frank para a família MGC-CS-CED, em que a ordem das distribuições componentes dispostas no modelo de misturas é importante para a identificação do modelo (ver Exemplo 4), foram estimados 17 modelos de mistura generalizados e 11 modelos de mistura tradicional. Para a comparação do ajuste dos modelos da família MGC-CS-CED e de mistura tradicional segundo os critérios de ajuste, a Tabela 4.5 mostra os cinco modelos estimados que obtiveram melhores ajustes segundo o critério BIC entre as 28 especificações analisadas. Os modelos de misturas generalizados foram os que apresentaram os melhores ajustes segundo o critério AIC, mas ficam atrás do modelo de mistura tradicional de componentes gama e log-normal, o melhor no critério BIC. O modelo de misturas tradicional de componentes log-logística e log-normal é o quarto melhor ajuste segundo o critério BIC. Segundo o critério AIC, esses dois modelos apresentaram os nono e décimo quarto melhores ajustes. Os modelos de misturas generalizados atribuíram probabilidades aproximadamente 62% e 28% às suas primeira e segunda parcelas, o que foi bem próximo dos pesos atribuídos às componentes do modelo de mistura tradicional. Na família MGC-CS-CED a primeira componente não é deformada, participando no modelo de forma proporcional, e o valor de dependência da cópula de Frank estimado acima de $\tau = 0,80$ para as três especificações MGC-CS-CED da tabela mostra que esses modelos selecionam para a mistura predominantemente os valores da região à direita da segunda componente. Também, os valores dos parâmetros das distribuições componentes de mesma especificação são bastante próximos. O modelo SNMIX ajustado a esses dados em Lin *et al.* (2007) tem sete parâmetros e obteve AIC e BIC iguais a $97,84$ e $122,35$, valores entre o sexto e sétimo melhores AIC e vigésimo segundo e vigésimo terceiro melhores BIC em comparação com as 28 especificações ajustadas. A Figura 4.14 mostra as funções densidade dos modelos estimados da tabela e também do modelo SNMIX. Apesar do ajuste geral dos modelos de mistura generalizados, mistura tradicional e SNMIX serem equivalentes, representando a densidade em torno de cada moda, os modelos de misturas generalizados exibem maior flexibilidade para representar os dados. Também, se o ajuste do modelo de mistura generalizado com parâmetros fixos para os indivíduos apresentou melhor ajuste que o modelo SNMIX, que flexibiliza os pesos das componentes para cada indivíduo da amostra com a inclusão de variáveis auxiliares latentes na estimação Bayesiana, então é esperado que ajustes ainda melhores do modelo de misturas generalizado possam ser obtidos com esses métodos, diversificando os parâmetros da cópula ou da mistura.

Tempo de erupções do gêiser The Old Faithful

O conjunto de dados consiste de 272 tempos de duração em minutos de erupções do gêiser The Old Faithful no Parque Nacional de Yellowstone, Wyoming, EUA. Vários conjuntos de dados similares sobre os tempos de erupção e também de espera entre erupções do gêiser são analisados na literatura e aqui analisamos a versão de Hardle (1990).

Os tempos aparentam ser uma mistura de duas componentes assimétricas. Azzalini e Bowman (1990) analisam os tempos de duração das erupções sucessivas do gêiser segundo um processo Markoviano de segunda ordem. Assim como Lin *et al.* (2007), que ajustaram o modelo SNMIX, iremos ajustar

Tabela 4.5: *Critérios de ajuste AIC e BIC e parâmetros dos modelos de mistura de percentis por conjuntos simples CS-CED e mistura tradicional estimados para o conjunto de dados de enzimas. São mostrados os cinco modelos de melhor ajuste pelo critério BIC.*

Modelo	AIC	BIC	Mistura	τ	θ	Componente 1		Componente 2	
MGC-CS-CED-Frank-Gam-Wei	96,60	117,61	0,62 (0,56;0,68)	0,82 (0,54;0,89)	20,78 (6,49;35,08)	4,83 (3,75;5,91)	0,04 (0,03;0,05)	0,86 (0,75;0,98)	1,46 (1,18;1,73)
MGC-CS-CED-Frank-Llog-Wei	96,75	117,76	0,63 (0,57;0,70)	0,86 (0,39;0,92)	26,04 (3,97;48,11)	0,18 (0,17;0,20)	3,59 (3,05;4,13)	0,86 (0,74;0,97)	1,44 (1,17;1,71)
MGC-CS-CED-Frank-Gam-Gam	96,84	117,85	0,62 (0,56;0,68)	0,80 (0,38;0,88)	17,70 (3,95;31,45)	4,82 (3,75;5,90)	0,04 (0,03;0,05)	2,40 (1,23;3,56)	0,34 (0,20;0,48)
MTrad-Gam-Lnor	98,73	116,23	0,62 (0,56;0,68)			4,90 (3,80;5,99)	0,04 (0,03;0,05)	0,23 (0,16;0,30)	0,33 (0,27;0,38)
MTrad Llog-Lnor	100,32	117,83	0,63 (0,57;0,69)			0,18 (0,17;0,20)	3,65 (3,13;4,18)	0,24 (0,17;0,31)	0,32 (0,27;0,37)

modelos de mistura generalizados à distribuição não condicional dos tempos de erupção do gêiser; isto é, não iremos a distribuição condicional dada pelo processo Markoviano.

Para os dados foram estimados os modelos de mistura generalizados da família MGC-CS-O (Exemplo 4) e de mistura tradicional com as combinações de distribuições componentes citadas. Como no primeiro conjunto de dados, temos 11 modelos para cada uma das propostas de misturas consideradas, a generalizada e a tradicional. Na Tabela 4.6 são mostradas as estimativas dos modelos de misturas que apresentaram melhor AIC e BIC entre todas as combinações estimadas e também, para comparação, o modelo de misturas tradicional de melhor AIC, de componentes log-normal e Weibull. Devido à diferença do valor da função de log-verossimilhança entre os modelos de mistura generalizados e de mistura tradicional e que os modelos pertencentes a cada grupo têm mesmo número de parâmetros, os critérios AIC e BIC estabelecem a mesma relação de ordem de qualidade de ajuste entre os modelos. O modelo de misturas tradicional obteve o décimo segundo melhor ajuste segundo os critérios AIC e BIC. Os modelos de mistura generalizados atribuíram cerca de 36% e 64% de probabilidade para as suas primeira e segunda parcelas. Esses valores foram próximos aos atribuídos às componentes estimadas para o modelo de misturas tradicional. A dependência alta negativa da cópula para os modelos de misturas generalizados da tabela mostra que nesses modelos a mistura toma predominantemente valores das regiões à direita da primeira componente e à esquerda da segunda componente. Também, as componentes log-normais estimadas por esses modelos obtiveram valores praticamente iguais. A partir dos parâmetros divulgados para o modelo SNMIX em Lin *et al.* (2007) e a amostra dos 272 tempos, temos para esse modelo de sete parâmetros o critério AIC de $529,14$, o que equivale a um ajuste entre o décimo primeiro e décimo segundo melhores ajustes, e BIC de $554,38$, entre os décimo primeiro e décimo segundo melhores ajustes, enquanto que o modelo de misturas generalizado com duas componentes normais ajustado obteve AIC de $528,35$ e BIC de $549,98$, o oitavo melhor ajuste. A Figura 4.15 mostra as funções densidade para os modelos ajustados e também o modelo SNMIX. Os modelos de misturas generalizados puderam representar muito bem as formas de assimetria contidas nos dados com bimodalidade. O modelo de misturas generalizado com especificação de componentes normais também acompanhou muito bem os efeitos de assimetria, exibindo melhor ajuste que o modelo equivalente SNMIX. Também nesses dados parece ser efetivo para melhorar o ajuste dos modelos de mistura generalizados com o uso de métodos de estimação da análise Bayesiana que flexibilizam os parâmetros da mistura e da cópula, eventualmente com covariáveis que houver disponíveis.

%

Tabela 4.6: *Critérios de ajuste AIC e BIC e parâmetros dos modelos generalizados da família MGC-CS-O e mistura tradicional estimados para o conjunto de dados de tempo de duração de erupções do gêiser The Old Faithful. São mostrados os modelos de melhor ajuste pelos critérios AIC e BIC e também o modelo de misturas tradicional de melhor ajuste para comparação.*

Modelo	AIC	BIC	Mistura	τ	θ	Componente 1		Componente 2	
MGC-CS-O-Frank-Lnor-Lnor	526,46	548,09	0,36 (0,30;0,41)	-0,85 (-0,90;-0,74)	-25,39 (-37,11;-13,66)	0,48 (0,43;0,53)	0,22 (0,18;0,26)	1,54 (1,51;1,56)	0,15 (0,13;0,17)
MGC-CS-O-Frank-Llog-Lnor	526,49	548,12	0,36 (0,30;0,42)	-0,84 (-0,89;-0,71)	-22,70 (-33,35;-12,05)	1,64 (1,57;1,71)	8,13 (6,30;9,96)	1,54 (1,52;1,56)	0,14 (0,12;0,16)
MGC-CS-O-Frank-Lnor-Gam	526,60	548,23	0,35 (0,30;0,41)	-0,85 (-0,89;-0,73)	-24,08 (-35,24;-12,92)	0,48 (0,43;0,53)	0,22 (0,17;0,26)	49,46 (36,48;62,43)	0,09 (0,07;0,12)
MTrad-Lnor-Wei	537,14	555,17	0,34 (0,29;0,40)			0,69 (0,67;0,72)	0,11 (0,09;0,13)	4,46 (4,40;4,52)	11,78 (10,31;13,26)

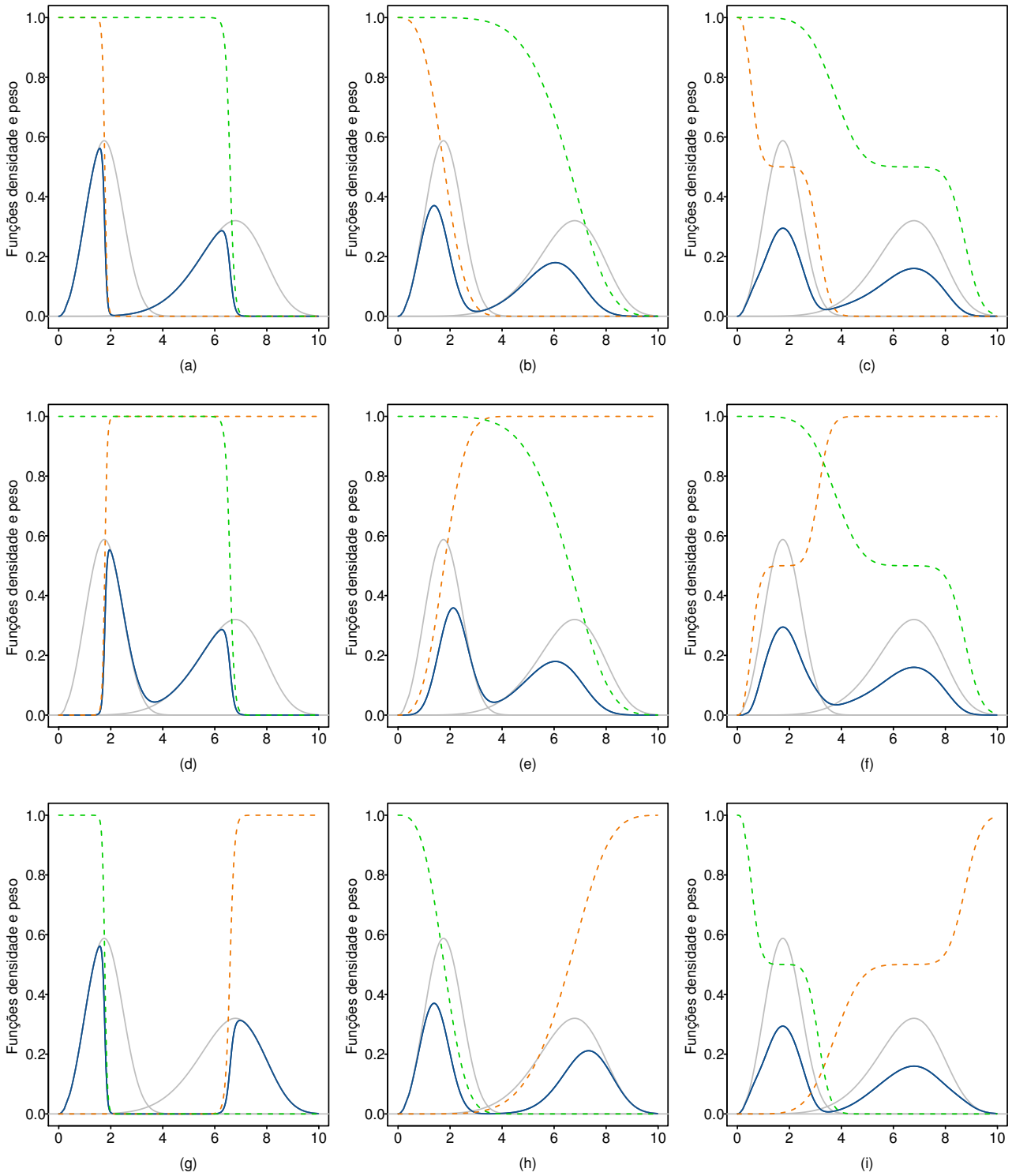


Figura 4.1: *De (a) a (c) o modelo MGC-CR-EE, do Exemplo 7, com cópula de Frank e distribuições componentes* $Weibull(2,0;3,0)$ *e* $Weibull(7,0;6,0)$*. A dependência da cópula é dada por* τ *de Kendall igual a* $-0,8$ *em (a),* 0 *em (b) e* $0,8$ *em (c). A função densidade do modelo de mistura é mostrada em linha contínua azul; a função peso da primeira componente em linha tracejada alaranjada e da segunda componente em linha tracejada verde; e as duas componentes em linha cinza. De (d) a (f), o modelo MGC-CR-ED, do Exemplo 8, com cópula de Frank, mesmos valores de dependência que (a) a (c), e distribuições componentes* $Weibull(2,0;3,0)$ *e* $Weibull(7,0;6,0)$*. De (g) a (i) o modelo MGC-CR-ED para o mesmo Exemplo 8 ilustrado em (d) a (f), porém trocando-se a ordem das componentes no modelo.*

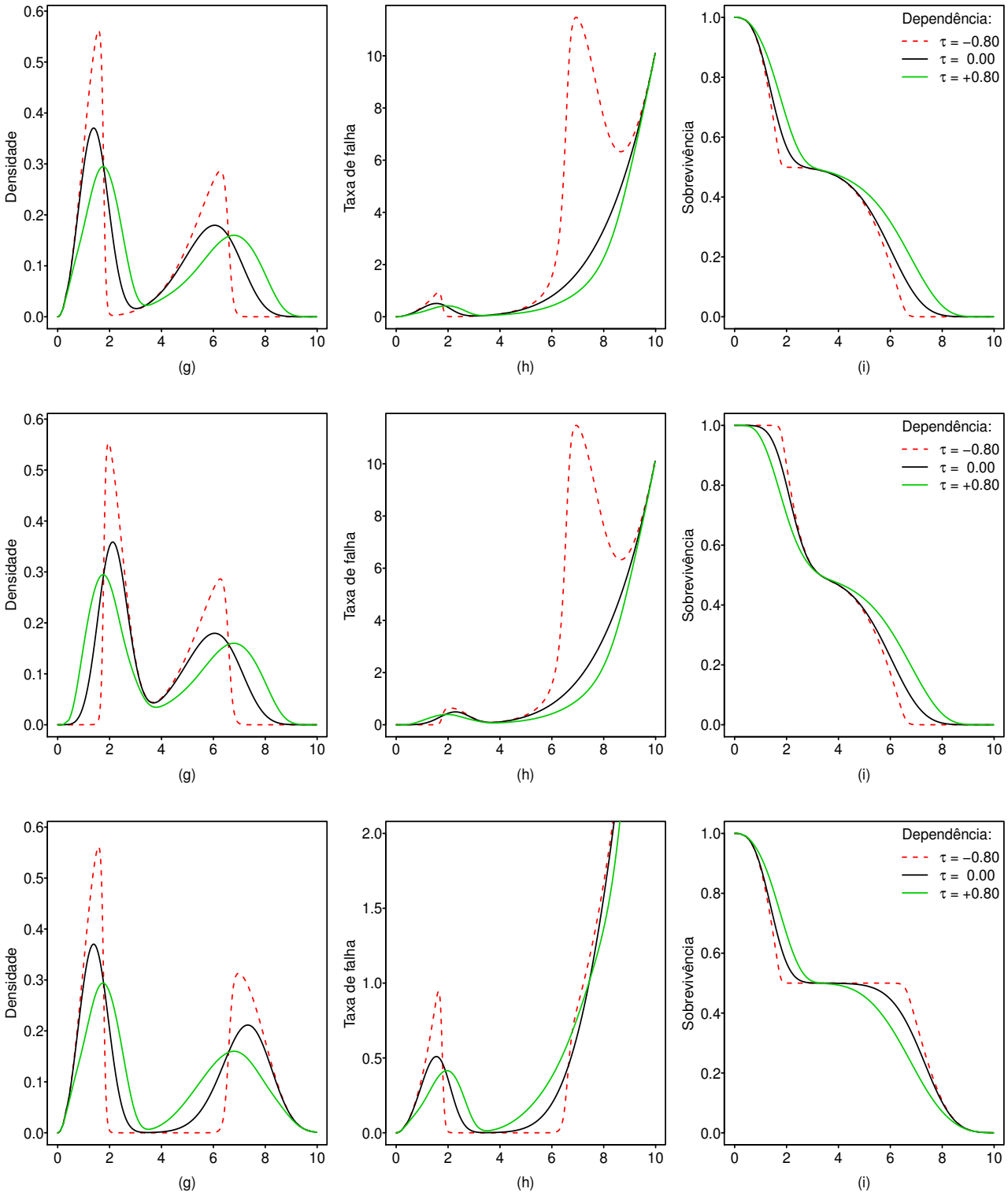


Figura 4.2: *Funções densidade (a), taxa de falha (b) e de sobrevivência (c) para o modelo MGC-CR-EE, do Exemplo 7, com cópula de Frank e distribuições componentes* $Weibull(2,0;3,0)$ *e* $Weibull(7,0;6,0)$. *De (d) a (f) as funções para o modelo MGC-CR-ED, do Exemplo 8, com cópula de Frank e distribuições componentes* $Weibull(2,0;3,0)$ *e* $Weibull(7,0;6,0)$. *De (g) a (i), as funções para o modelo MGC-CR-ED, do mesmo Exemplo 8 ilustrado em (d) a (f), porém trocando-se a ordem das componentes no modelo.*

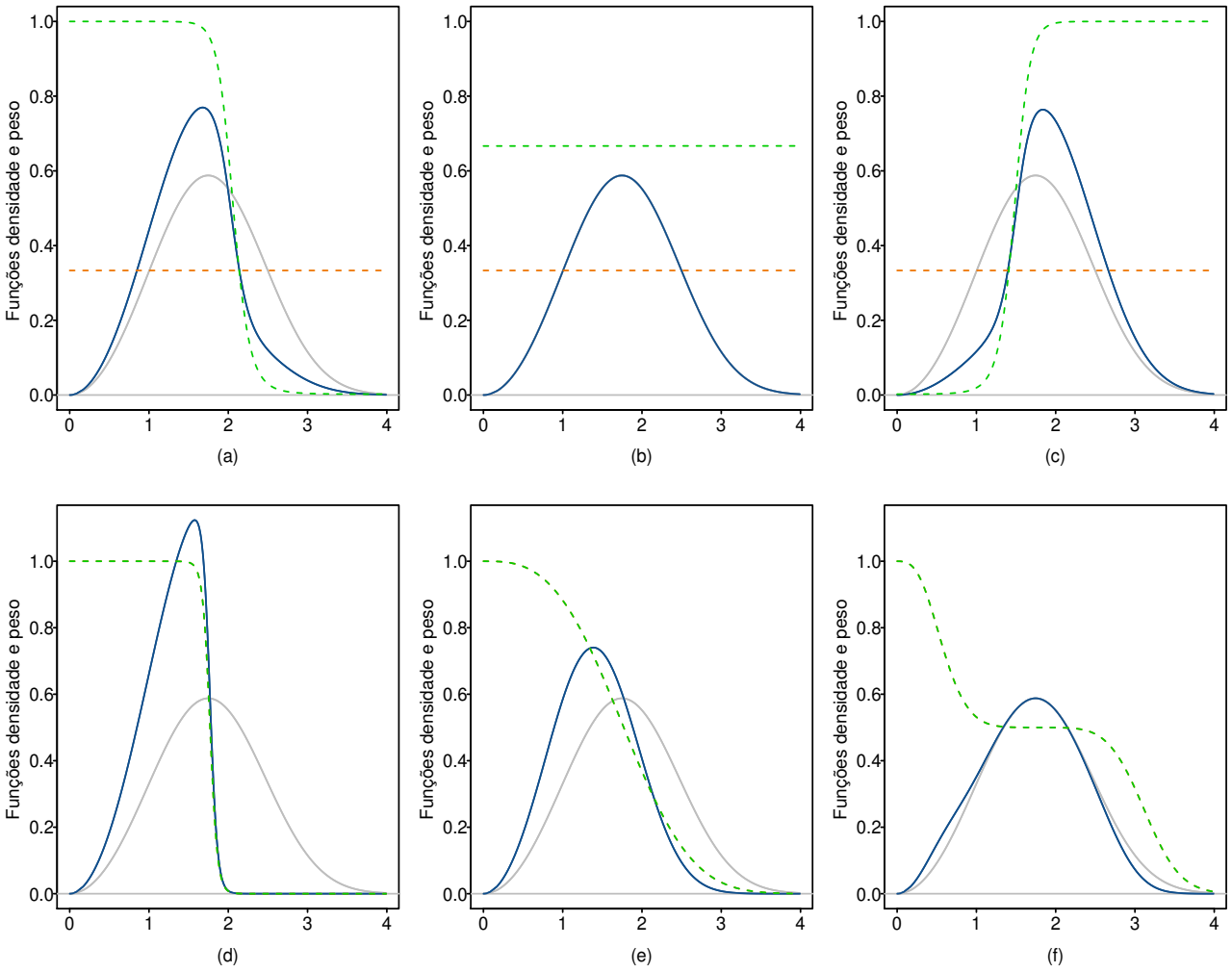


Figura 4.3: *De (a) a (c) o modelo MGC-CS-EA, do Exemplo 9, com cópula de Frank, parâmetro da mistura* $p = 1/3$ *e ambas distribuições componentes* $Weibull(2, 0; 3, 0)$*. A dependência da cópula é dada por* τ *de Kendall igual a* $-0,8$ *em (a),* 0 *em (b) e* $0,8$ *em (c). A função densidade do modelo de mistura é mostrada em linha contínua azul; as funções peso da primeira componente em linha tracejada alaranjada e da segunda componente em linha tracejada verde; e as componentes iguais em linha cinza. De (d) a (f), o modelo MGC-CR-EE, do Exemplo 10, com cópula de Frank, mesmos valores de dependência que (a) a (c), e distribuições componentes* $Weibull(2, 0; 3, 0)$*. As funções peso de ambas componentes são superpostas em linha tracejada verde.*

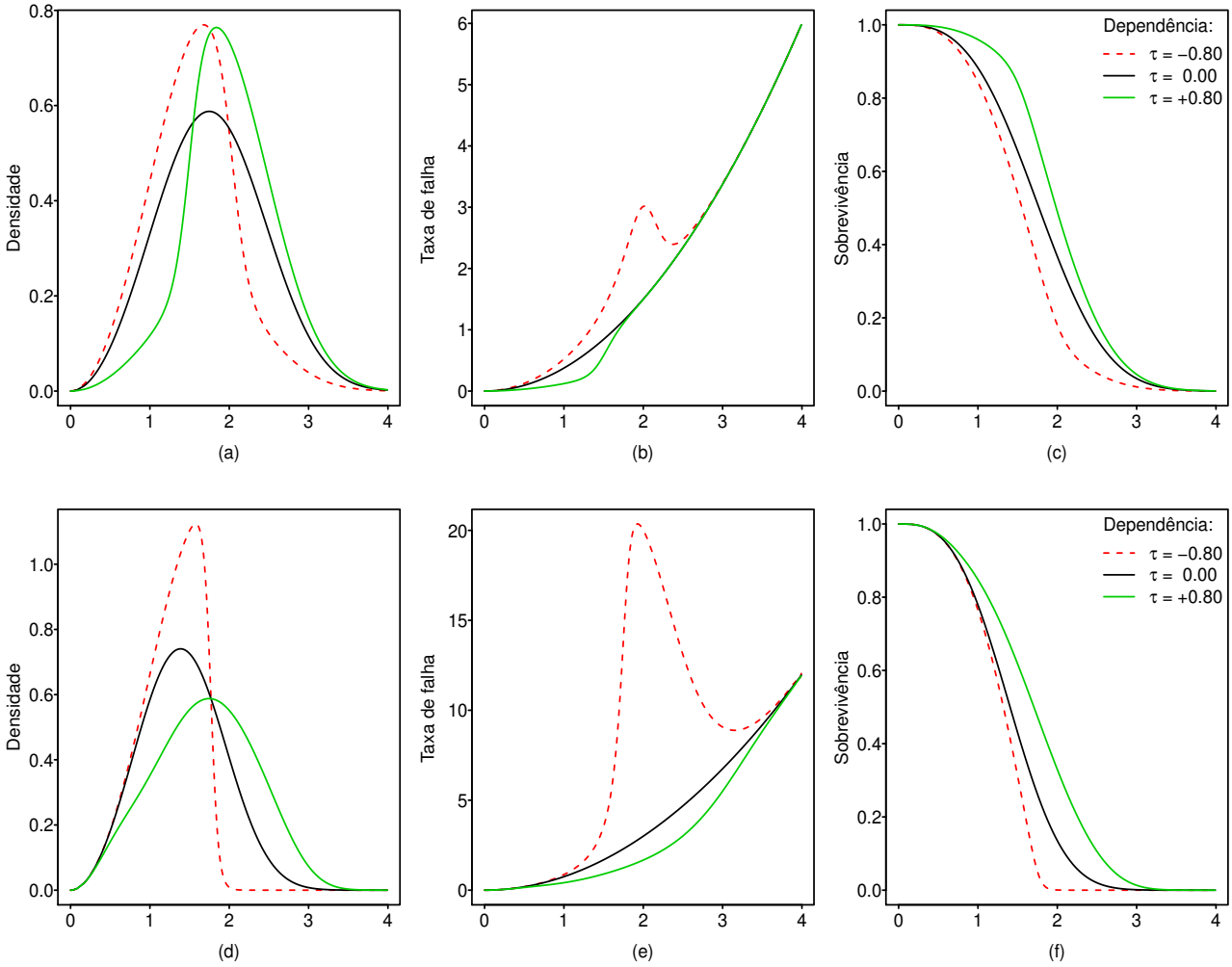


Figura 4.4: *Funções densidade (a), taxa de falha (b) e de sobrevivência (c) para o modelo MGC-CS-EA, do Exemplo 9, com cópula de Frank, parâmetro da mistura* $p = 1/3$ *e ambas distribuições componentes* $Weibull(2, 0; 3, 0)$. *De (d) a (f) as funções do modelo MGC-CR-EE, do Exemplo 10, com cópula de Frank e ambas distribuições componentes* $Weibull(2, 0; 3, 0)$.

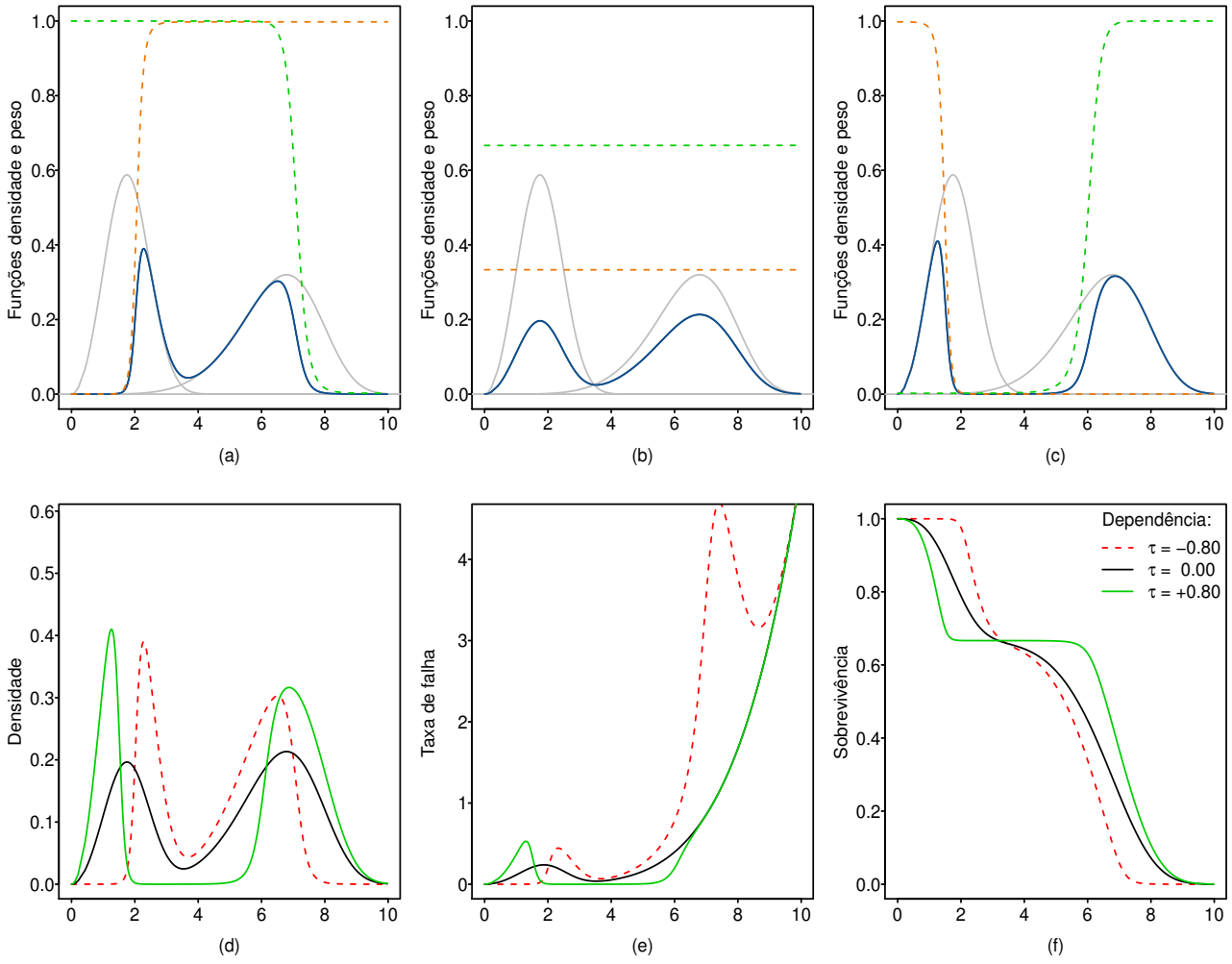


Figura 4.5: *De (a) a (c) o modelo MGC-CS-O, do Exemplo 11, com cópula de Frank, parâmetro da mistura* $p = 1/3$ *e distribuições componentes* $Weibull(2,0;3,0)$ *e* $Weibull(7,0;6,0)$*. A dependência da cópula é dada por* τ *de Kendall igual a* $-0,8$ *em (a),* 0 *em (b) e* $0,8$ *em (c). A função densidade do modelo de mistura é mostrada em linha contínua azul; a função peso da primeira componente em linha tracejada alaranjada e da segunda componente em linha tracejada verde; e as duas componentes em linha cinza. As funções densidade, taxa de falha e de sobrevivência para o modelo são mostradas em (d), (e) e (f), respectivamente.*

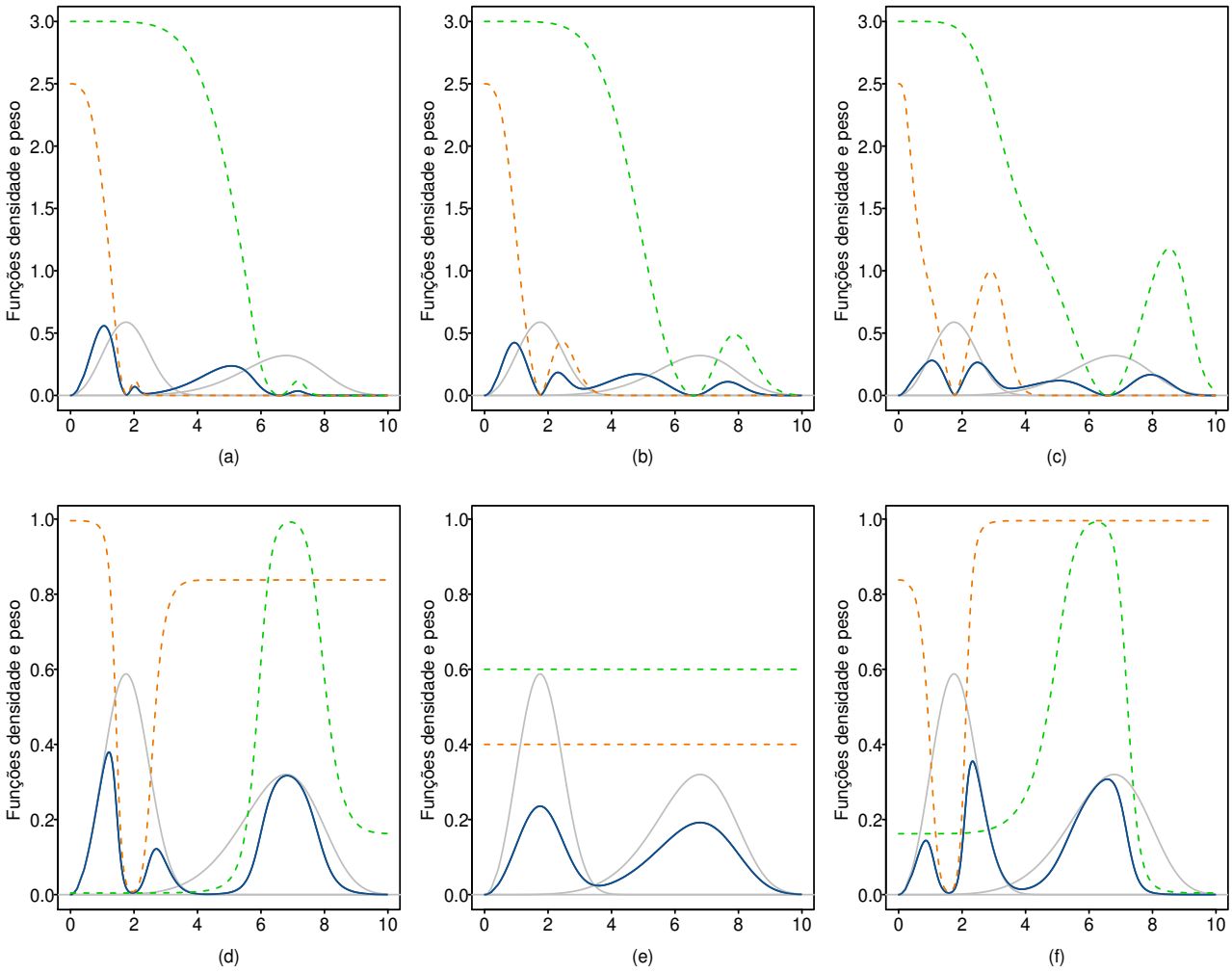


Figura 4.6: *De (a) a (c) o modelo MGC-CR-CAE, do Exemplo 12, com cópula de Frank, parâmetros da mistura* $k_1 = 2,5$ *e* $k_2 = 3,0$ *e distribuições componentes* $Weibull(2,0;3,0)$ *e* $Weibull(7,0;6,0)$*. A dependência da cópula é dada por* τ *de Kendall igual a* $-0,8$ *em (a),* 0 *em (b) e* $0,8$ *em (c). A função densidade do modelo de mistura é mostrada em linha contínua azul; a função peso da primeira componente em linha tracejada alaranjada e da segunda componente em linha tracejada verde; e as duas componentes em cinza. De (d) a (f), o modelo MGC-3CS-CC, do Exemplo 13, com cópula de Frank, mesmos valores de dependência que (a) a (c), parâmetros da mistura* $p = 0,1$ *e* $q = 0,7$ *e distribuições componentes* $Weibull(2,0;3,0)$ *e* $Weibull(7,0;6,0)$*.*

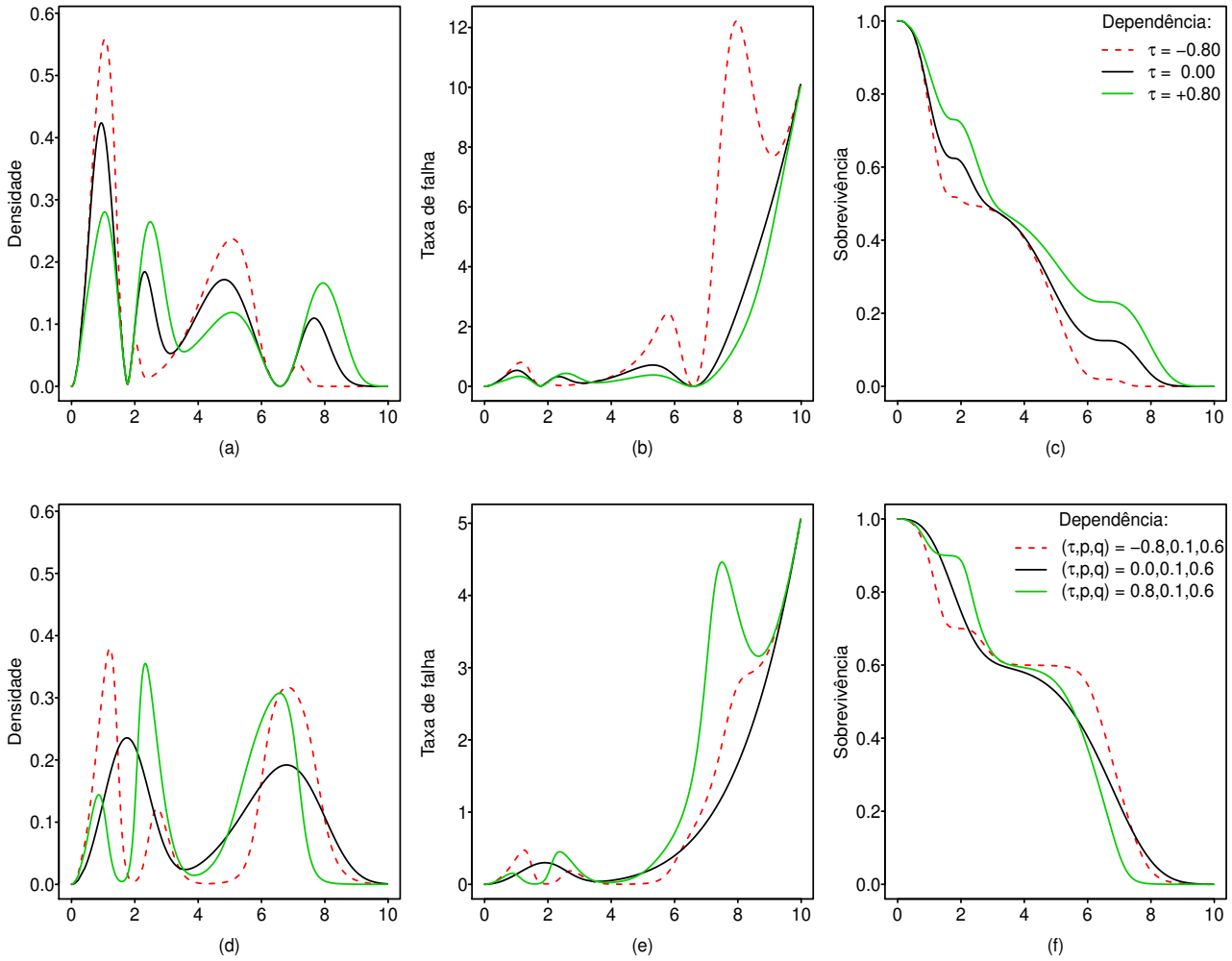


Figura 4.7: *Funções densidade (a), taxa de falha (b) e de sobrevivência (c) para o modelo MGC-CR-CAE, do Exemplo 12, com cópula de Frank, parâmetros da mistura* $k_1 = 2,5$ *e* $k_2 = 3,0$ *e distribuições componentes* $Weibull(2,0;3,0)$ *e* $Weibull(7,0;6,0)$. *De (d) a (f) as funções para o modelo MGC-3CS-CC, do Exemplo 13, com cópula de Frank, parâmetros da mistura* $p = 0,1$ *e* $q = 0,7$ *e distribuições componentes* $Weibull(2,0;3,0)$ *e* $Weibull(7,0;6,0)$.

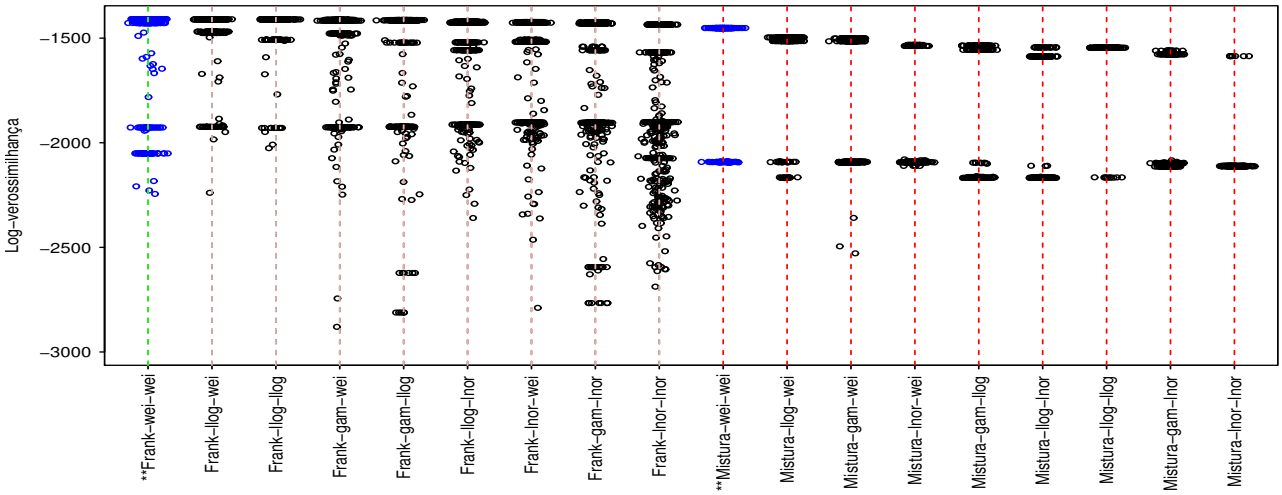


Figura 4.8: *Valores de máximo resultantes da otimização da função de log-verossimilhança de cada modelo proposto a partir de* 300 *a* 600 *valores iniciais, ajustando a uma amostra de tamanho* $n = 1.000$ *gerada do modelo MGC-CR-EE. Os modelos ajustados são apresentados em ordem decrescente de maior valor obtido para a função de log-verossimilhança. A especificação correta, dos dados simulados, é o primeiro modelo, os asteriscos indicam os modelos com distribuições componentes corretas.*

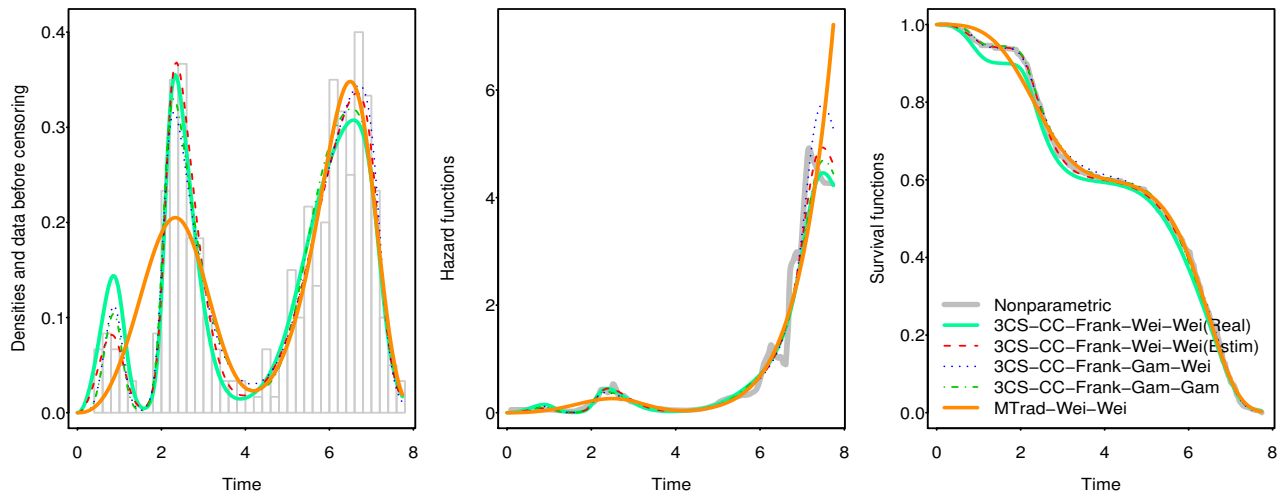


Figura 4.9: *Funções densidade, taxa de falha e sobrevivência para os modelos MGC-3CS-CC e mistura tradicional ajustados à amostra de tamanho n=300 do Caso 1.*

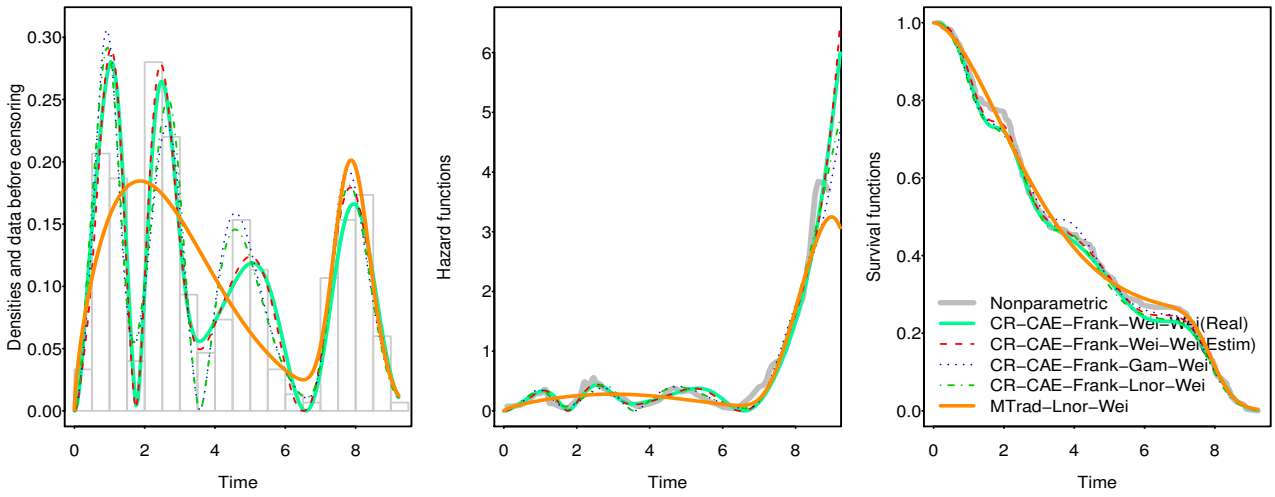


Figura 4.10: *Funções densidade, taxa de falha e sobrevivência para os modelos MGC-CR-CAE e mistura tradicional ajustados à amostra de tamanho n=300 do Caso 2.*

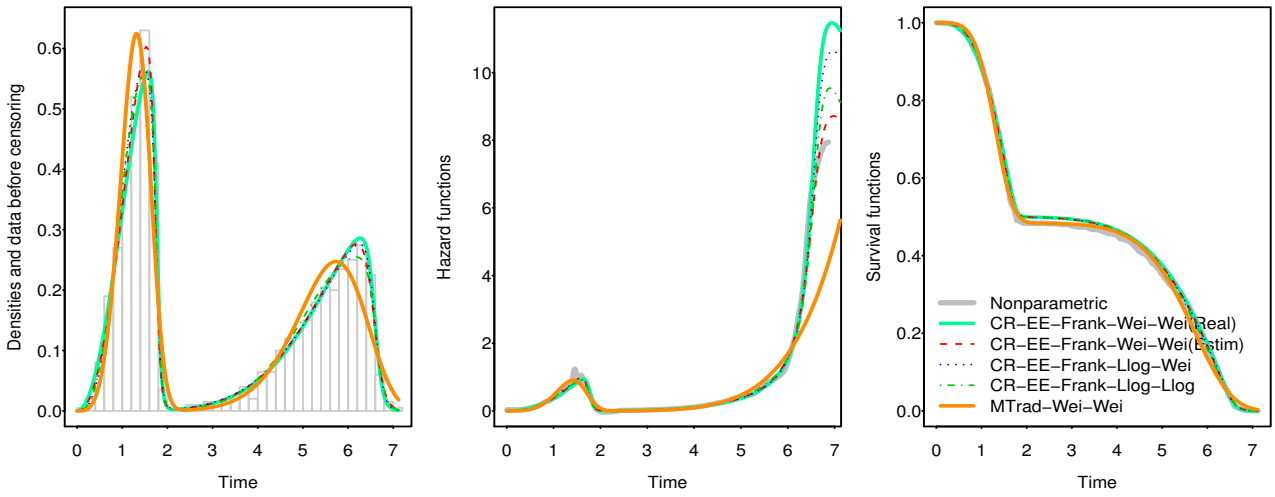


Figura 4.11: *Funções densidade, taxa de falha e sobrevivência para os modelos MGC-CR-EE e mistura tradicional ajustados à amostra de tamanho n=300 do Caso 3.*

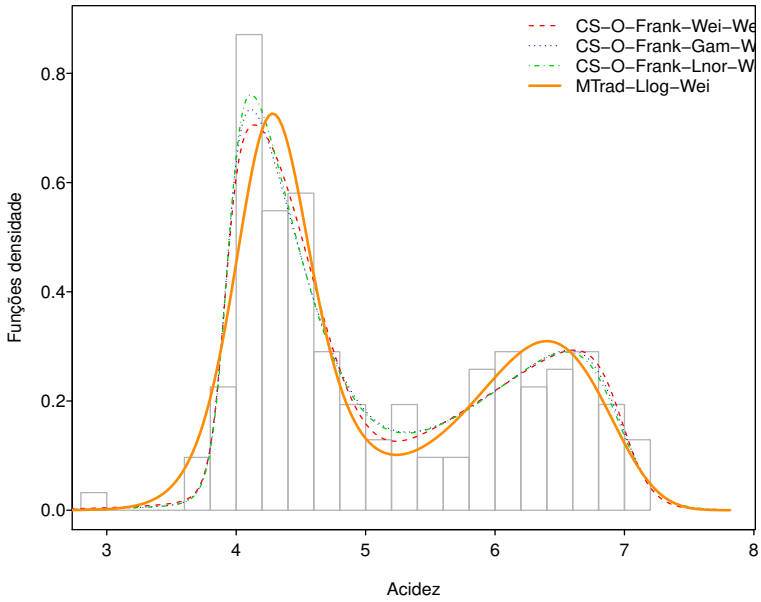


Figura 4.12: *Comparação das funções densidade dos modelos MGC-CS-O e de mistura tradicional estimados para o conjunto de dados de acidez da água em lagos.*

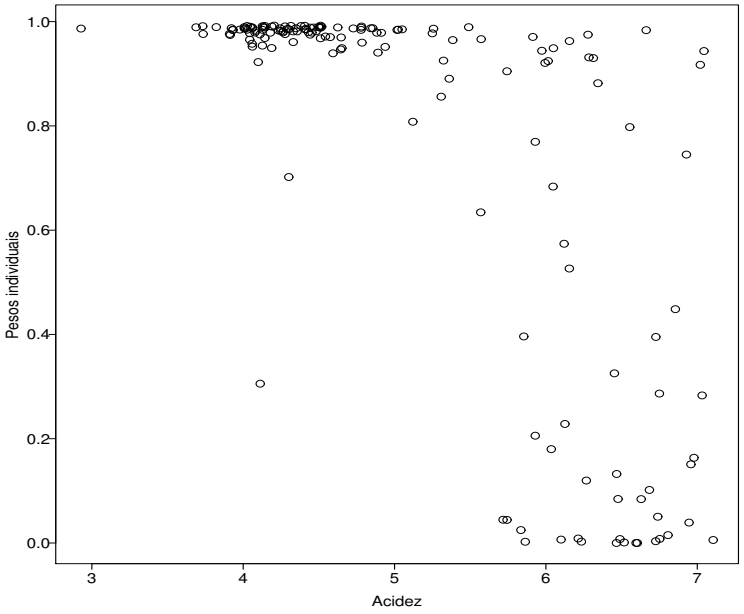


Figura 4.13: *Relação das observações de acidez dos lagos com escores estimados como pesos individuais na análise Bayesiana de Crawford et al. (1992) utilizando as covariáveis disponíveis no conjunto de dados e uma função logística estimada.*

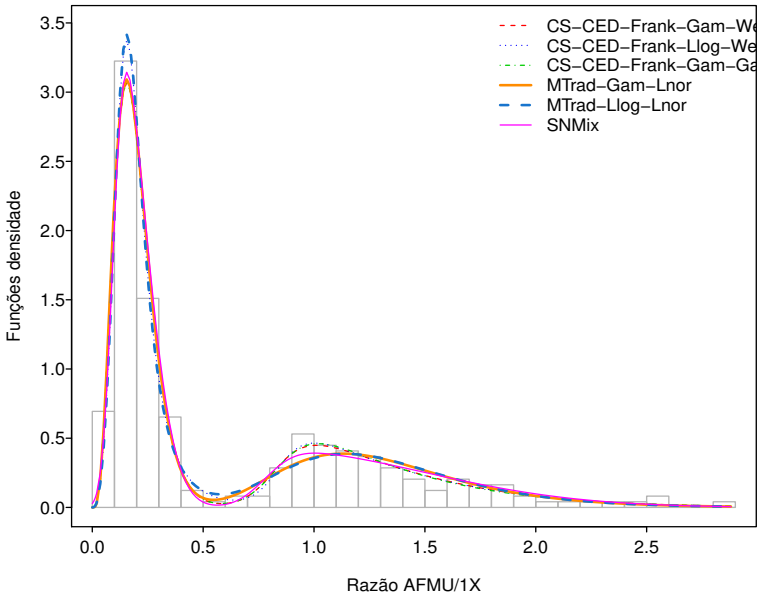


Figura 4.14: *Comparação das funções densidade dos modelos generalizados da família MGC-CS-CED e de mistura tradicional estimados para o conjunto de dados de enzimas.*

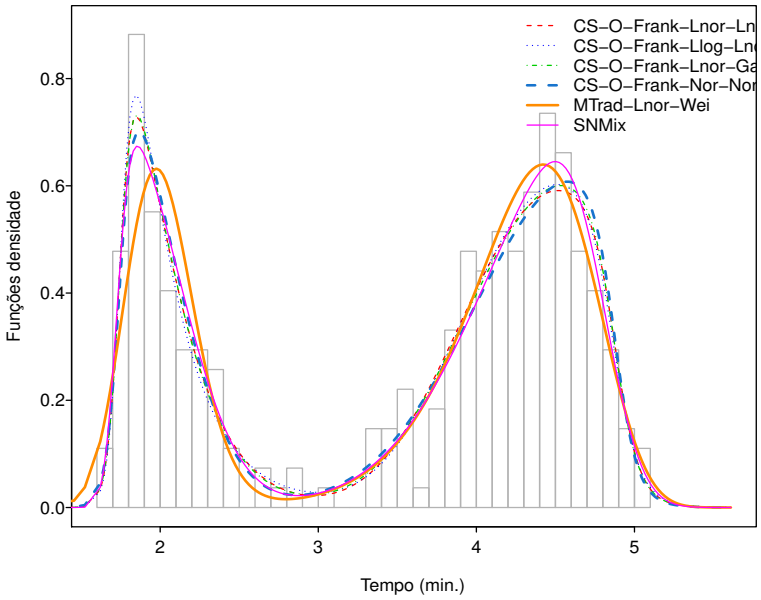


Figura 4.15: *Comparação das funções densidade dos modelos generalizados da família MGC-CS-O e mistura tradicional estimados para o conjunto de dados de tempo de duração de erupções do gêiser The Old Faithful. Também é mostrado o modelo de misturas de distribuições skew normal estimado em Lin et al. (2007).*

Capítulo 5

Conclusões

Nessa tese propomos duas metodologias para a generalização de modelos utilizando cópulas. Esses modelos são o modelo de riscos múltiplos e o modelo de misturas de distribuições.

Os modelos de riscos múltiplos independentes são conhecidos como uma ferramenta flexível para a construção de funções de taxa de falha. Sua formulação é baseada em tempos latentes de falha. Verificamos que o uso de funções de cópulas para modelar a dependência dos tempos latentes aumenta consideravelmente essa flexibilidade. Nessa primeira proposta de generalização de modelos utilizando cópulas, os modelos de riscos múltiplos com dependência possibilitam construir uma família rica de funções de taxa de falha com formas de banheira, multimodais e contendo efeitos locais. O modelo mostra muito bom ajuste aos dados simulados e de duração de desemprego, representando os efeitos resultantes da concorrência entre os riscos. Nessa proposta não foi possível fazer inferências sobre os tempos latentes, devido ao problema de identificação resultante da falta de informação sobre as causas de falha. No entanto, o modelo proposto permite a inclusão, convenientemente, de restrições na dependência da cópula como, por exemplo, negativa, positiva ou presente nas caudas, e também permite o exame da associação direta entre covariáveis e o comportamento dos tempos latentes.

Analisamos uma extensão dessa primeira proposta de metodologia, eliminando a restrição de suporte positivo das variáveis latentes. As qualidades de ajuste do modelo de riscos múltiplos com dependência são válidas para distribuições com suporte nos reais e permitem a geração de uma família rica de distribuições univariadas, contendo assimetrias e múltiplas modas. Nessa extensão, mostramos que os modelos podem aproximar densidades da família de distribuições generalizadas geradas da distribuição beta (Alexander *et al.*, 2011; Cordeiro *et al.*, 2012). Os modelos de riscos múltiplos com dependência estendidos ao suporte real mostraram bom ajuste a três conjuntos de dados da literatura. A riqueza de representações desses modelos pode ser aumentada utilizando funções de copula de dois parâmetros ou aumentando o número de variáveis latentes.

Na segunda proposta de modelagem dessa tese, analisamos o modelo de mistura de distribuições generalizado utilizando cópulas. O modelo proposto tem outros modelos da literatura como casos particulares, que são o modelo de mistura tradicional, o modelo de fração de cura e o modelo de riscos múltiplos dependente e independente, discutidos na primeira parte dessa tese. A participação das distribuições componentes na mistura proposta é dada por pesos que variam no suporte da distribuição, flexibilizando a capacidade de representação do modelo. Propomos algumas famílias de modelos que derivam da definição do modelo generalizado. Nessas famílias, os parâmetros da cópula e da mistura são utilizados para definir as características de assimetria e ponderar com maior ou menor peso determinadas regiões das distribuições componentes para compor o modelo de mistura. Resultam dessas especificações modelos bastante flexíveis, capazes de representar multimodalidade e assimetria dos dados. Essas características foram analisadas segundo algumas formas possíveis de representação das funções densidade, sobrevivência e taxa de falha, estas duas no caso de distribuições de variáveis assumindo valores nos reais positivos.

O modelo de mistura generalizado por cópulas foi ajustado a conjuntos de dados simulados para algumas das especificações propostas. O modelo proposto mostrou o melhor ajuste em relação aos modelos de mistura tradicional em geral, com maior flexibilidade para se ajustar aos dados. Na análise de dados empíricos da literatura o modelo mostrou muito bom ajuste a dados bimodais contendo assimetrias, apresentando ajustes melhores que o modelo de misturas tradicional e também melhores que os modelos de misturas com componentes normais assimétricas (*skew normal*) de Lin *et al.* (2007) segundo os critérios AIC, BIC e pela comparação dos gráficos das funções densidade estimadas com os histogramas das observações.

5.1 Sugestões para Pesquisas Futuras

Uma direção de investigação importante para o modelo de riscos múltiplos com dependência é a inclusão de covariáveis. Segundo Klein e Moeschberger (2003), os modelos de vida acelerados têm uso limitado devido ao número de distribuições que eles podem assumir. Um resultado possível dessa abordagem é uma proposta de modelo de vida paramétrico baseado no modelo de riscos múltiplos com dependência. O modelo tem uma estrutura latente que permite associar covariáveis diretamente ao comportamento dos tempos latentes de falha, o que deve permitir ao modelo uma boa flexibilidade de ajuste a dados em que há heterogeneidade. Essa associação de covariáveis também pode ser feita em relação ao parâmetro de dependência da cópula.

Algumas direções envolvendo os modelos de mistura generalizado por cópulas podem ser exploradas na sequência desse trabalho, entre elas o desenvolvimento de resultados assintóticos e formalização das propriedades sugeridas, a inclusão de covariáveis, análise de identificabilidade, análise de outros modelos da literatura que são seus casos particulares e também a análise de outras especificações do modelo dadas por diferentes funções de cópulas e de funções de seleção utilizadas.

Muitas vezes a análise de dados requer a consideração de covariáveis que são disponíveis nos conjuntos de dados. A inclusão de covariáveis no modelo de misturas pode ser feita de forma que seus níveis sejam associados aos efeitos de assimetria (em tipo ou intensidade), à predominância de determinadas regiões do suporte da distribuição, à região das caudas da distribuição, à participação das transformações das componentes no modelo final, conforme a parametrização do modelo de mistura.

Outra direção possível também envolvendo covariáveis é a exploração da relação das formas do modelo com covariáveis disponíveis nos dados visando resultados de agrupamento, por meio de análise Bayesiana com variáveis auxiliares (Tanner e Wong, 1987). Como exemplo, nos modelos de misturas de Lin *et al.* (2007) os parâmetros dos pesos das distribuições componentes são ajustados como diferentes para cada indivíduo nos conjuntos de dados de enzimas e tempo de erupções. Nesses casos, em que os modelos de misturas utilizando cópulas com parâmetros constantes obtiveram melhores ajustes, é uma possibilidade explorar o potencial do modelo flexibilizando os parâmetros de dependência e de misturas em função de covariáveis por métodos Bayesianos.

As especificações consideradas para o modelo de mistura de distribuições utilizando cópulas nesse trabalho foram escolhidas com o objetivo de mostrar o potencial do método em produzir modelos de misturas capazes de ajustar a dados com efeitos de assimetrias e multimodalidade. No entanto, outras especificações precisam ser exploradas. Determinadas combinações de funções de cópulas e distribuições componentes podem produzir modelos com formas analíticas de menor complexidade ainda contendo as mesmas propriedades que as apresentadas nos casos selecionados.

Outro problema de interesse é a análise do potencial do modelo de mistura utilizando cópulas em incluir outros modelos da literatura como seus casos particulares. Como exemplo, o modelo assimétrico '*skew-symmetric*' de Azzalini (1986) e Henze (1985), dado por $\hat{f}(t) = 2G(at)f(t)$, em que G é uma função de distribuição, a é uma constante e $f(t)$ uma função densidade simétrica, pode ser analisado como caso particular de uma especificação do modelo de misturas utilizando cópulas, devido à sua

forma analítica semelhante.

Apêndice A

Polyhazard Models With Dependent Causes

Polyhazard Models with Dependent Causes

Rodrigo Tsai e Luiz Koodi Hotta

A aparecer no *Brazilian Journal of Probability and Statistics*.
http://www.imstat.org/bjps/

Accepted manuscript - 12/01/2012 - Brazilian Journal of Probability and Statistics

Polyhazard Models with Dependent Causes

Rodrigo Tsai[a,b] **and Luiz Koodi Hotta**[a]

[a] *State University of Campinas - UNICAMP*
[b] *Superior Court of Justice - STJ/Brazil*

Abstract. Polyhazard models constitute a flexible family for fitting lifetime data. The main advantages over single hazard models include the ability to represent hazard rate functions with unusual shapes and the ease of including covariates. The primary goal of this paper was to include dependence among the latent causes of failure by modeling dependence using copula functions. The choice of the copula function as well as the latent hazard functions results in a flexible class of survival functions that is able to represent hazard rate functions with unusual shapes, such as bathtub or multimodal curves, while also modeling local effects associated with competing risks. The model is applied to two sets of simulated data as well as to data representing the unemployment duration of a sample of socially insured German workers. Model identification and estimation are also discussed.

1 Introduction

Polyhazard models are a flexible family for fitting lifetime data. Their flexibility stems from the acknowledgment that there are latent causes of failure. There are many applied examples of these models in the literature. Kalbfleisch and Prentice (1980) proposed the poly-log-logistic model for log-logistic competing risks; Berger and Sun (1993) proposed the poly-Weibull model for Weibull competing risks; Louzada-Neto (1999) proposed a generalized polyhazard model which encompasses the poly-Weibull, poly-log-logistic and generalized-poly-gamma models; Kuo and Yang (2000) and Basu et al. (1999) used the poly-Weibull model to model masked-systems, in which the cause of failure may be unknown or partially known; Mazucheli et al. (2001) presented a Bayesian inference procedure for the polyhazard models with covariates; and Louzada-Neto et al. (2004) analyzed the identifiability of the poly-Weibull model. The main advantage of polyhazard models compared to single hazard models is the flexibility to represent hazard rate functions with unusual shapes.

In the applications cited above, the latent causes of failure are independent. In this paper, we extend the independent polyhazard models to en-

AMS 2000 subject classifications. Primary: 62N99, 62H99, 62P05.
Keywords and phrases. Polyhazard Models, Copula, Competing Risks.

compass dependence modeled by copula functions. The model is general enough to allow for various forms of dependence and also for any marginal distributions for the latent times. The proposed models are able to generate much more flexible risk functions than the independent polyhazard models, including features such as bathtub shape, multimodality and local effects.

The literature also mentions another approach for constructing flexible hazard functions that is not pursued here. In this approach, the authors generalize known distributions. See, for instance, Pham and Lai (2007) and Nadarajah et al. (2011). The method proposed in the present paper, however, is more general. For instance, each of these distributions can be used as a marginal distribution for the latent causes.

The polyhazard model with dependence is proposed in Section 2. In Section 3 identification and estimation of the model through maximum likelihood method is discussed. Another option would be to use a Bayesian approach; however, this is tangential to the purpose of this paper, as is model estimation, and thus is not discussed in detail. In Section 4, we present applications of simulated data and of data on unemployment duration of German women who are part of the socially secured workforce. General remarks are presented in Section 5.

2 The polyhazard model with dependence

Consider that we observe n units of observations, each one subject to $k \geq 2$ competing latent causes of failure. Let the lifetime related to the jth latent cause of the ith unit of observation, X_{ij}, have a density $f_j(\cdot;\Gamma_j)$, which are considered as known except for the unknown set of parameters Γ_j. Denote the survival and hazard functions by $S_j(\cdot;\Gamma_j)$ and $\lambda_j(\cdot;\Gamma_j)$, respectively. Only $X_i = \min\{X_{ij}, j = 1, \cdots, k\}$ is observed for each unit of observation. Thus, considering the independence among risks, namely, among the failure times X_{ij}, $j = 1, \cdots, k$, the overall survival function of X_i, denoted by $S(t;\Upsilon)$, where $\Upsilon = (\Gamma_1, \cdots, \Gamma_k)$, is given for any $i = 1, \cdots, n$ by the product of marginal survival functions, i.e.

$$\begin{aligned} S(t;\Upsilon) &= P_\Upsilon[X_i > t] \\ &= P_\Upsilon[X_{i1} > t, \cdots, X_{ik} > t] \\ &= \prod_{j=1}^{k} S_j(t;\Gamma_j), \end{aligned} \tag{2.1}$$

and the hazard function of X_i, $\lambda(t;\Upsilon)$, is given by the sum of the marginal hazards, because

$$\begin{aligned}\lambda(t;\Upsilon) &= \frac{-\frac{d}{dt}\prod_{j=1}^{k} S_j(t;\Gamma_j)}{\prod_{j=1}^{k} S_j(t;\Gamma_j)} \\ &= \sum_{j=1}^{k} \lambda_j(t;\Gamma_j). \end{aligned} \tag{2.2}$$

An example of an application of the independent polyhazard model is given in Mazucheli et al. (2001) where they estimate the poly-Weibull model with covariates using a Bayesian approach. In this paper, we model the failure time X_i with $k = 2$ competing risks, allowing for dependence between the risks. Henceforth, we use the notation for $k = 2$ for simplicity, but the notation for $k > 2$ can be easily generalized. Denoting by $H(\cdot,\cdot;\Upsilon)$ the joint distribution function and by $\bar{H}(\cdot,\cdot;\Upsilon)$ the joint survival function of the latent variables X_{i1} and X_{i2}, we can write the survival function of X_i as

$$\begin{aligned} S(t;\Upsilon) &= P_{\Upsilon}[X_{i1} > t, X_{i2} > t] \\ &= \bar{H}(t,t;\Upsilon). \end{aligned} \tag{2.3}$$

To model the joint survival function $\bar{H}$, considering dependence between the latent variables, we propose the use of copula functions. An m-dimensional copula function may be defined as a cumulative distribution function whose marginal distributions are uniform over $[0,1]$ and whose support is the $[0,1]^m$ hypercube. Copula functions have been extensively studied in the multivariate modeling literature, especially when the use of the multivariate normal distribution is questionable. An important feature of the copula approach is the possibility of modeling the dependence and the marginal behavior of the related variates separately, thus making the copula a very convenient alternative in the case of multivariate modeling. Some references for copulas include the textbooks of Nelsen (2006), Joe (1997) and Cherubini et al. (2004) as well as the paper of Trivedi and Zimmer (2005).

Let $F_1(\cdot;\Gamma_1)$ and $F_2(\cdot;\Gamma_2)$ be the distribution functions of X_{i1} and X_{i2}, respectively. It follows from Sklar's theorem that there is always a copula function C^* such that we can write $H(t_1,t_2;\Upsilon) = C^*(F_1(t_1;\Gamma_1), F_2(t_2;\Gamma_2))$ and that, C^* is unique if the marginal distributions F_1 and F_2 are continuous. C^* is then called a copula function because it couples the marginal distributions F_1 and F_2 to their joint distribution H. It is possible to represent the joint survival function directly by $\bar{H}(t_1,t_2;\Upsilon) = P[X_1 > t_1, X_2 > t_2;\Upsilon] = \tilde{C}(S_1(t_1;\Gamma_1), S_2(t_2;\Gamma_2))$, where $\tilde{C}(u,v) = u+v-1+C^*(1-u,1-v)$ is

4

also a copula. On the other hand, for any copula C, $C(S_1(t_1;\Gamma_1), S_2(t_2;\Gamma_2))$ is a survival distribution function. Therefore, we can also model the survival function S directly by a copula function C as in Kaishev et al. (2007). This is also the approach adopted here because it is generally easier to work analytically with this representation. Then for the survival function of the polyhazard model with dependence given by a copula function C with dependence parameter θ and $\Upsilon = (\theta, \Gamma_1, \Gamma_2)$, we can write

$$\begin{aligned} S(t;\Upsilon) &= \bar{H}(t,t;\Upsilon) \\ &= C_\theta(S_1(t;\Gamma_1), S_2(t;\Gamma_2)), \end{aligned} \tag{2.4}$$

where S_1 and S_2 are, in this paper and in almost all practical applications, continuous marginal survival functions. The copula C in (2.4) is called the survival copula; in this paper, we refer to it as the copula function. Notice that the right (left) tail dependence for the latent survival times is equal to the left (right) tail dependence of copula C of (2.4). From the survival function (2.4), it follows that the probability density and hazard rate functions for the polyhazard model with dependence are obtained in the usual fashion, that is

$$f(t;\Upsilon) = -\frac{d}{dt}S(t;\Upsilon) \quad \text{and} \quad h(t;\Upsilon) = \frac{f(t;\Upsilon)}{S(t;\Upsilon)}. \tag{2.5}$$

The proposed model is a generalization of the independent polyhazard model in that we allow for dependence while at the same time modeling the marginal behavior of the latent risks. For each combination of copula and marginal survival functions employed, we have another model that allows for the construction of a rich family of competing risks latent models. For instance, in the following sections, we will work with exponential, log-logistic, log-normal, Gamma and Weibull distributions for the latent failure causes and Clayton, Gumbel and Frank copula functions. However, we could work with any distribution and any copula function. The symmetrized Joe Clayton (SJC) copula is not used in the applications, although it is used as an example in some parts of the paper. These copula functions were selected because they have been widely used in the literature and have different types of dependence. The Frank copula, with parameter $\theta \in (-\infty, +\infty)$, is a symmetric Archimedean copula with Kendall's $\tau \in (-1, 1)$ and Spearman's $\rho \in (-1, 1)$, and with lower and upper tail dependence λ_L and λ_U equal to zero. While it can generate distributions with strong dependence in the center, the dependence in the tails is always small. Thus, in the tails, the hazard function of the competing risks model will be approximately equal to the sum of the marginal hazard functions. For the Clayton copula, the parameter $\theta \in$

$(0, +\infty)$, $\tau = \theta/(\theta + 2) \in [0, 1)$, $\rho \in [0, 1)$, $\lambda_U = 2^{-1/\theta} \in (0, 1)$, and $\lambda_L = 0$. For the Gumbel copula, the parameter $\theta \in [1, +\infty)$, $\tau = (\theta - 1)/\theta \in [0, 1)$, $\rho \in [0, 1)$, $\lambda_U = 0$, and $\lambda_L = 2 - 2^{1/\theta} \in [0, 1)$. For the SJC copula λ_L and $\lambda_U \in [0, 1)$. These features must be taken into consideration when selecting the copula function (see Trivedi and Zimmer (2005) for more properties). In the above discussion, we always referred to the dependence between the latent variables.

As an example of a specification of the polyhazard model with dependence, consider the Frank copula and Weibull latent failure times such that $X_{ij} \sim Weibull(\mu_j; \beta_j)$, $j = 1, 2$. This model will be referred to as Frank-Weibull-Weibull, where the first name stand for the copula function and the last two names denote the latent distributions. According to the notation of the proposed model, its parameters can be denoted by $\Upsilon = (\theta, \Gamma_1, \Gamma_2)$, where $\Gamma_1 = (\mu_1; \beta_1)$ and $\Gamma_2 = (\mu_2; \beta_2)$. The overall survival function of X_i is given by

$$S(t; \Upsilon) = -\frac{1}{\theta} \log\left(1 - \frac{(1 - e^{-\theta e^{-(t/\mu_1)^{\beta_1}}})(1 - e^{-\theta e^{-(t/\mu_2)^{\beta_2}}})}{(1 - e^{-\theta})}\right), \qquad (2.6)$$

and the probability density of X_i by

$$f(t; \Upsilon) = \frac{(1 - e^{-\theta S_2(t)})e^{-\theta S_1(t)} f_1(t) + (1 - e^{-\theta S_1(t)})e^{-\theta S_2(t)} f_2(t)}{(1 - e^{-\theta}) - (1 - e^{-\theta S_1(t)})(1 - e^{-\theta S_2(t)})}, \qquad (2.7)$$

where f_1 and f_2 are the density functions of X_{i1} and X_{i2}, respectively. Figure 1 illustrates some possible shapes for the distribution of X_i for the Frank-Weibull-Weibull specification, considering $X_{i1} \sim W(4; 0.9)$ and $X_{i2} \sim W(5; 3)$ and the dependence parameter varying in a range where the Kendall's τ ranges from -0.80 to 0.80. The figure shows that different shapes for the hazard rates can result, depending on the shapes of the marginal distributions and the dependence type. Figure 2 shows various hazard rate functions for other specifications of the model in which it is possible to notice local effects and bathtub and multimodal shapes. The two points in the figure denote the 99% and 99.9% quantiles for each specification and the dependence parameter between the latent variables is the Kendall's τ, except for the SJC copula where they denote the lower and upper tail dependence. Henceforth we use the acronyms Lnor, Llog, Exp, Wei, Gam and Indep for the log-normal, log-logistic, exponential, Weibull and gamma distributions and the independence copula, respectively, when referring to a specification of the polyhazard model. For instance, Clayton-Llog-Wei refers to a polyhazard model with the Clayton copula and log-logistic and Weibull latent variables.

6

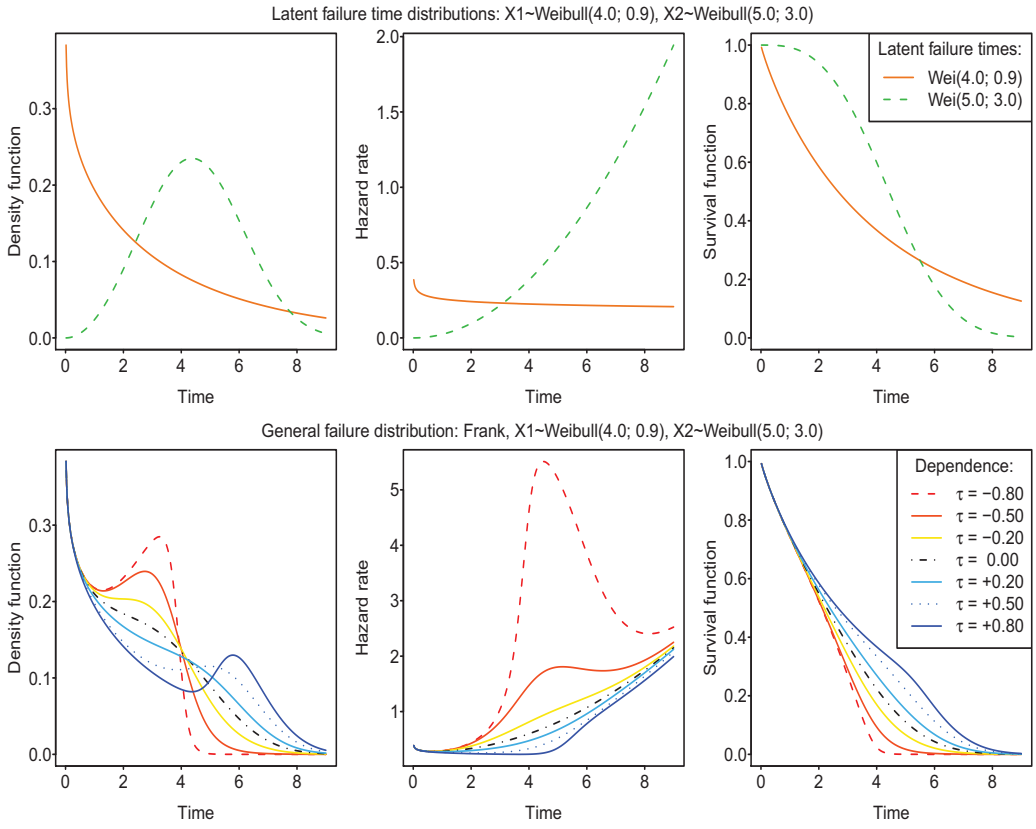


Figure 1 *Examples of density, hazard and survival functions for the single risk Weibull model and polyhazard model with Weibull marginals and dependence through Frank copula and Weibull marginals.*

3 Model Identification and Estimation

Some models are clearly non-identifiable. Consider, for instance, the model Indep-Exp-Exp whose overall hazard function is constant, say $\lambda > 0$, where the latent hazard function can be any non-negative constant, say λ_1 and λ_2, such that $\lambda = \lambda_1 + \lambda_2$. A less trivial non-identifiable model is the dependent polyhazard model Gumbel-Wei-Wei. The Gumbel copula function is given by

$$C(u, v) = \exp[-\{(-\log u)^\theta + (-\log v)^\theta\}^{\frac{1}{\theta}}], \quad u, v \in [0, 1].$$

Therefore, by (2.4), considering Weibull margins with parameters functions (λ_1, β_1) and (λ_2, β_2), the overall survival function is given by

$$\begin{aligned} S(t) &= C(S_1(t), S_2(t)) \\ &= \exp[-\{\lambda_1^\theta t^{\theta\beta_1} + \lambda_2^\theta t^{\theta\beta_2}\}^{\frac{1}{\theta}}], \end{aligned}$$

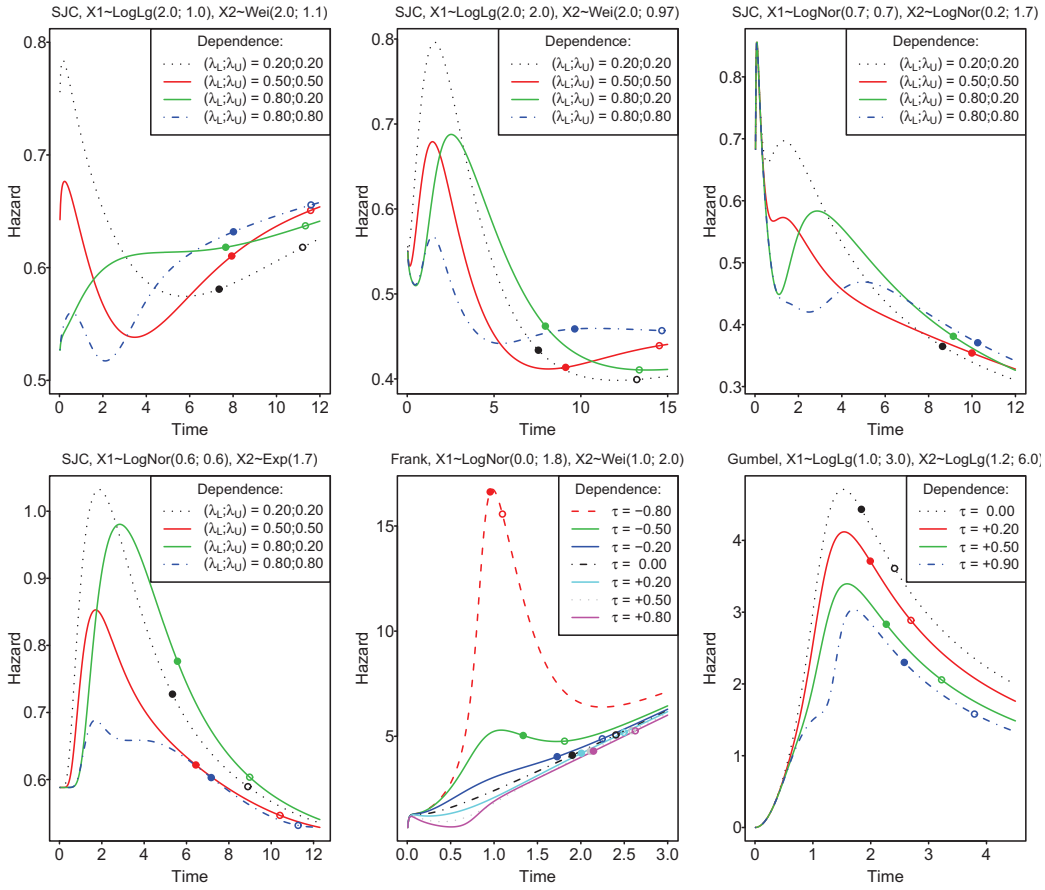


Figure 2 *Examples of hazard rate functions for the polyhazard model with dependence.*

showing that the model is not identifiable when $\beta_1 = \beta_2 = \beta$ for which any triple $(\lambda_1', \lambda_2', \theta')$ satisfying $(\lambda_1'^{\theta'} + \lambda_2'^{\theta'})^{1/\theta'} = (\lambda_1^{\theta} + \lambda_2^{\theta})^{1/\theta}$ can generate the same model. The same non-identification problem occur in the subclass of the Gumbel-Wei-Wei models: Gumbel-Exp-Exp, Indep-Exp-Exp, and Indep-Wei-Wei models. Another example of non identifiable model is the Clayton-Llog-Llog model when both marginal distributions have the same shape parameter and the dependence parameter equals 1. In general, we also have nonidentifiability when the distribution of one marginal latent variable is stochastically dominated by the other latent distribution and we use a copula with perfect positive dependence. Usually, it is not that easy to check whether a dependent polyhazard model is identifiable or not by analytical analysis. For this reason, identification of the other models given by combinations of the Clayton, Gumbel and Frank copulas with the exponential, log-logistic, log-normal, Gamma and Weibull latent cause dis-

8

tributions was conducted by two types of numerical analyses. In the first analysis, the identification of each specification of the model was analyzed by means of an optimization procedure that searched over a region of parametric space for different points representing equal density functions. The analysis covered $1,000$ points that were sampled uniformly in a hyperspace that was a Cartesian product of individual parameter sets that was considered wide enough to represent the parametric space for real situations. For the dependence parameter, we considered the Kendall's τ in $[-0.99, 0.99]$ for the Frank copula and in $[0.01, 0.99]$ for Clayton and Gumbel copula functions. For the latent variables parameters: exponential's scale in $[0.01; 4.00]$; gamma's form in $[0.01; 10]$ and scale $[0.01; 8]$; log-logistic's scale in $[0.01; 8]$ and form in $[0.01; 8]$; log-normal's location in $[-3; 3]$ and form in $[0.5; 3]$; and Weibull's scale in $[0.01; 10]$ and shape in $[0.01; 3]$. Then, for each of these $1,000$ points, its density function was evaluated in a grid of 301 points to serve as reference of a search of another point that could produce the same density function. Denote by $M(\Upsilon)$ the model under investigation, where Υ is its set of parameters in the parametric space E_Υ. For Υ_0, one of the $1,000$ arbitrarily chosen points in E_Υ, the overall density of $M(\Upsilon_0)$ was evaluated in a grid with 301 points, the $100(0.005 + 0.99\, i/300)\%$ quantiles, $i = 0, \cdots, 300$. The algorithm looked for a point in the parametric space that minimizes the objective function, $D(\Upsilon, \Upsilon_0)$, the sum of squared errors in which the errors were the differences between the density functions of $M(\Upsilon_0)$ and $M(\Upsilon)$ on the grid. For each Υ_0 the optimization step was repeated by 10 initial values, summing $10,000$ cases, so that for the i-th initial value denote by $\Upsilon_{0,i}$ the value located by the algorithm. After the optimization analysis the cases where $D(\Upsilon_0, \Upsilon_{0,i}) < 10^{-16}$ and $d(\Upsilon_0, \Upsilon_{0,i}) = \sum_{j=1}^{p}[(\upsilon_{0,i,j} - \upsilon_{0,j})/\upsilon_{0,j}]^2 > 10^{-10}$ were considered as indication of non-identifiability, where $\Upsilon_0 = (\upsilon_{0,1}, \cdots, \upsilon_{0,p})$ and $\Upsilon_{0,i} = (\upsilon_{0,i,1}, \cdots, \upsilon_{0,i,p})$. Every case satisfying these conditions were analysed individually. The procedure detected the special cases of the Gumbel-Wei-Wei, Indep-Wei-Wei and Clayton-Llog-Llog models mentioned before as non-identifiable. In the second analysis, in the applications with the real dataset and with the simulated data, we used different initial points, numbering approximately 200, for the optimization of the likelihood function in all cases. Except for a few cases of local maxima, the convergences were at the same values. In this study of convergence, we used more simulated datasets than the two presented in the illustration section. The analysis showed that, except for the cases mentioned previously there was strong evidence of identification for all other specifications. A different point, estimability, is discussed more in the following paragraphs. An identifiable model does not ensure easy parameter

estimation. For instance, when the overall hazard function is dominated by the first latent cause, it is very difficult to estimate the second latent cause, except for large samples.

In the traditional competing risks literature, when the cause of failure is known, there is another type of discussion of identification. See, for instance, Cox (1972) and Tsiatis (1975). In this classical problem, a competing risks model is identifiable if the joint survival function can be calculated or identified by the simple knowledge of the overall survival distribution. Tsiatis (1975) found that, for a model with dependent risks, it is possible to find a set of independent risks that produces the same joint survival distribution. It follows that, unless restrictions are imposed on the behavior of the competing risks, this type of identification is not possible. Some papers exhibit results in this direction. Heckman and Honoré (1989) use a function that is similar to a copula based on covariates to overcome, nonparametrically, the identification problem. Carriere (1994) relates the marginal crude probabilities to the net probabilities using copula functions when there is dependence among the risks. Zheng and Klein (1995) show that the identification of the marginal distributions is possible if the copula function is fixed.

The polyhazard model can be seen as a competing risks model with missing values for the cause. Because less information is available, identification of the equivalent competing risks model is necessary but not sufficient for the identification of the polyhazard model. However, even when we have this type of non-identification in polyhazard models, we can still use these models to model lifetime data and thus benefit from the good characteristics of these models.

The model parameters are estimated by maximum likelihood method. Considering a random sample X_i, $i = 1, \cdots, n$, with random right censoring in which δ_i is the failure indicator variable and t_i the minimum value of the failure and censoring times, it follows from (2.4) and (2.5) that the likelihood is given by

$$L(\Upsilon) = \prod_{i=1}^{n} f(t_i; \Upsilon)^{\delta_i} S(t_i; \Upsilon)^{1-\delta_i},$$

where Υ denotes the parameters for the copula function and the marginal distributions. The algorithms were written in R and the log-likelihood functions were implemented in C for fast computation. The optimization used the Nelder-Mead algorithm; in all applications, we tested for several initial parameter values to check for possible problems of local maxima and identification. Except for the issue of local maxima observed in the estimation of the copula specifications, we did not find convergence problems in several applications using both empirical and simulated data.

10

The analysis of the Hessian matrix shows that for some specifications, a large number of observations is necessary to have a small variance of the estimator of the copula parameter. This is especially important when the difference between the polyhazard model with dependence and the independent polyhazard model lies in a region with small probability. This is expected because a large number of overall observations are needed to have a reasonable number of observations in the region of small probability.

4 Illustrations

This section presents illustrations for simulated data, using two models and for the real data on the duration of female unemployment in Germany. For each dataset, all models given by the combinations of the exponential, log-logistic, log-normal, Gamma and Weibull distributions for the latent failure causes and the Clayton, Gumbel, Frank and Independent copulas were fitted, except for the Indep-Exp-Exp and Gumbel-Exp-Exp models, which were not identifiable. The exponential, log-logistic, log-normal, Gamma and Weibull distributions, which are single risk models, were also fitted. Because there are many polyhazard models, we only present the fitting of some of these models. These include all single risk factor models and those polyhazard models selected according to the AIC criterion: the best specification for each copula function and for each dataset for models with AIC comparable with that of the best model. In the simulations, we also included for each copula the model with the right marginal specification. We consider datasets with and without censored observations.

4.1 Simulated Data

The first dataset is a random sample of size $N = 5,000$ from a Frank-Lnor-Wei model. The parameter of the Frank copula is given by $\theta = -5.74$, which gives Kendall's τ equal to -0.50. The Frank copula has both tail dependencies equal to zero. For the latent marginal distribution, we used log-normal ($\mu_1 = 0.6$; $\sigma_1 = 1.8$) and Weibull ($\mu_2 = 2.0$; $\beta_2 = 4.0$). A large sample size is necessary for this model to have sufficient observations in the right extreme tail. A random censoring mechanism was applied with uniform distribution $U(0; a\, x_{(n)})$, where $x_{(n)}$ is the maximum of the simulated latent values and $a = 5.3$. This resulted in 20% of the observations censored, while 43.1% of the observed data came from the first latent cause and 36.9% from the second cause. The upper panel of Figure 3 presents a graph where on the Y-axis, we have the cause of failure (1 for the first cause, 2 for the second cause and 3 if it is censored), and on the X-axis, we have the minimum of

11

the two latent failure times. We plotted only a sample of 500 observations to be able to visualize the points. We observe that almost all the smallest values came from the first latent cause, while for the large values, we have an inversion, although not as dominant as for small values. Table 1 presents the estimates of some single risk models and for the polyhazard models selected by the Akaike criterion. The Akaike criterion was calculated as $AIC = -2\,L(\hat{\Upsilon}) + 2k$, where k is the number of parameters and $L(\hat{\Upsilon})$ is the log-likelihood function evaluated at the maximum likelihood estimate. The parameters for the marginal distributions are as follows: exponential (scale); Weibull(form; scale); gamma(form; scale); log-logistic(scale; form) and log-normal(location; scale). The confidence intervals for the estimates are exhibited in parentheses and were calculated numerically from the Fisher information. In this example, the polyhazard models offered a better fit in terms of the AIC and in terms of adjustment to the nonparametric estimation of the density, hazard and survival functions relative to the single risk models. The first 4 models selected by the AIC criterion are Frank copula model (from a total of 63 models tested, 15 are Frank Copula model). In this simulated dataset, the selection of the right copula was likely facilitated due to the large sample size. Moreover, the Frank copula has no tail dependence and was generated with a negative Kendall' τ coefficient, while both the Gumbell and Clayton copulas have tail dependence and positive Kendall' τ coefficients. Table 1 presents the estimation of the fitted models. We also included the first four best models selected by the AIC criteiron, all of which are Frank copula model. Observe that when the lognormal distribution is selected for the model, its estimates are not far from the true marginal distributions, even when the fitted copula is wrong or when the other marginal distribution is specified incorrectly.

Figure 4 presents the theoretical values of the density, hazard and survival functions and their estimates using single risk models, polyhazard models selected by AIC (Frank-Lnor-Wei) and using a nonparametric method. The nonparametric estimate of the survival function is the Kaplan-Meier survival function estimate smoothed by the R-program Loess method. To estimate the hazard function, the derivatives are numerically computed from the smoothed survival function and the Loess filter was again applied to the numerical derivatives. The smoothing parameter was selected empirically for each case. The estimation can depend strongly on the parameter, especially in the extremes. The nonparametric and the polyhazard function methods provide good estimates, while the single risk models are not able to fit the data. This first illustration clearly demonstrates the greater flexibility of the polyhazard models compared to the single risk models. Figure 5 presents

12

the comparison of the fit of some polyhazard models. The estimates of the function density and survival function by all the polyhazard models selected by AIC criterion are close to the true functions. However, only the models with the Frank copula estimate the hazard function well for the entire period. The other specifications fail to fit the theoretical and nonparametric estimates of the hazard function in the right tail.

We used the same dataset to fit the eight copula models of Table 1 without censoring and with 10% and 30% of the observations censored. Considering the cases of 10% and 20% of censoring we have a total of 64 estimates of the risk parameters. Comparing with the estimates found without censoring the maximum relative difference was 6% for the point estimates and 23% for their standard deviations. These values were equal to 11% and 23% for the 16 estimates of the Kendall's τ and their standard deviations. The standard deviations were estimated using the delta method. For the 30% censoring the maximum realtive difference in the 32 estimates was 98%, and 69% for the point estimates and their standard deviations, respectively. These differences, however, is smaller when we considered that the second last differences were equal to 32%, and 47%.

The same exercise was repeated with sample sizes N equal to 100, 250, 500, $1,000$, $2,000$ and $10,000$ without censoring and with 20% of censored observations. For every case, the single risk model yielded a bad fit. The AIC selected a dependent copula over the independent copula only when the sample size was larger or equal to $N = 1,000$. This is somewhat expected because the main difference between both models occurs in the extreme right tail. The hazard function has a change in the curvature around time 2.15 and another change around time 2.5. To detect this change in the curvatures, it is necessary to have some observations in this region. Thus, it is not surprising that even when we simulated a sample as large as $2,000$, the estimated hazard function was not accurate at the extreme because, without censoring, the probability of observing a failure larger than 2.5 is 0.0039. That is the main reason why in this first example, we used a large sample. The estimation of the probability density and survival functions require fewer observations. For instance, Figure 6 presents the estimation of the same Frank-Lnor-Wei model with sample size equal to 500, $1,000$, $2,000$ and $5,000$, without censoring. All the estimates are close to the theoretical values.

The second example is a random sample of size $N = 1,000$ from a model with Clayton copula with parameter $\theta = 18$ (Kendall's $\tau = 0.90$, $\lambda_U = 0.96$ and $\lambda_L = 0$) and log-logistic$(13.0; 1.0)$ and log-normal$(3.0; 0.5)$ latent marginals. More than half (61.6%) of the observed data were ob-

13

Table 1 *Simulation 1. True model: Frank copula $\theta = -5.74$ (Kendall's $\tau = -0.50$) with log-normal$(0.6; 1.8)$ and Weibull$(2.0; 4.0)$ marginals and sample size equal to $5,000$. Single risk models and models selected by AIC criterion: best polyhazard models, best marginals configuration for each copula and the copula model for the right marginal configuration*

Model	AIC	τ	θ	Marg. Distrib. 1		Marg. Distrib. 2	
Frank-Lnor-Wei	7417.90	-0.51 (-0.62;-0.34)	-5.90 (-8.46;-3.33)	0.64 (0.53;0.76)	1.83 (1.74;1.93)	2.03 (1.90;2.15)	3.96 (3.35;4.57)
Frank-Lnor-Gam	7418.84	-0.66 (-0.71;-0.60)	-9.79 (-11.65;-7.93)	0.61 (0.50;0.72)	1.81 (1.73;1.90)	7.23 (5.98;8.47)	0.29 (0.23;0.34)
Frank-Lnor-Llog	7419.91	-0.67 (-0.72;-0.61)	-10.28 (-12.29;-8.27)	0.65 (0.53;0.76)	1.84 (1.75;1.93)	1.98 (1.93;2.04)	4.06 (3.65;4.46)
Frank-Lnor-Lnor	7420.83	-0.70 (-0.74;-0.65)	-11.40 (-13.46;-9.35)	0.58 (0.48;0.68)	1.79 (1.71;1.88)	0.72 (0.69;0.74)	0.43 (0.39;0.46)
Clayton-Lnor-Gam	7426.42	0.45 (0.36;0.53)	1.67 (1.11;2.22)	0.54 (0.45;0.63)	1.76 (1.69;1.84)	15.08 (12.55;17.62)	0.09 (0.08;0.11)
Clayton-Lnor-Wei	7427.26	0.58 (0.49;0.65)	2.80 (1.90;3.70)	0.76 (0.64;0.88)	1.90 (1.81;2.00)	1.45 (1.42;1.48)	3.18 (2.96;3.40)
Gumbel-Lnor-Wei	7427.31	0.39 (0.10;0.54)	1.65 (1.11;2.19)	0.60 (0.49;0.70)	1.80 (1.72;1.89)	1.53 (1.45;1.61)	3.54 (3.14;3.93)
Indep-Lnor-Wei	7432.81			0.67 (0.57;0.78)	1.86 (1.77;1.94)	1.69 (1.67;1.72)	4.27 (4.06;4.48)
Weibull	8605.86			1.22 (1.20;1.25)	1.51 (1.47;1.54)		
Gamma	8919.70			1.58 (1.52;1.64)	0.72 (0.69;0.76)		
Exponential	9407.66			1.19 (1.16;1.23)			
Log-logistic	9825.31			0.95 (0.93;0.98)	1.77 (1.73;1.82)		
Log-normal	10243.41			-0.18 (-0.21;-0.15)	1.08 (1.06;1.10)		

tained from the first latent cause and 31.5% were obtained from the second cause. The censoring mechanism was the same as in the previous case with $a = 3$ producing 6.9% of censored observations. The lower panel of Figure 3 presents the same graph as in the first simulated dataset, also including only 500 observations. Almost all the smallest values came from the first latent cause and there is less mixture in the middle in comparison with the first example. Table 2 shows the estimates for the polyhazard models with the best fit for each copula according to the Akaike criterion and the models of single risk. Except for the independent copula, which was ranked only 19-th in terms of AIC, the other polyhazard models produced a fit close to the nonparametric hazard function estimate. Because the single risk model again produced a bad fit, in Figure 7, we present only the results for the polyhazard models of Table 2. In this example, it is observed that when one or both of the marginals are correctly specified, the parameter estimates of the correctly specified variables are very close to their true value. In this example, we also observed the same facts we observed in respect to the estimates of the marginal distributions and the effect of censoring.

Similarly to the first example, we fitted models with different sample sizes,

14

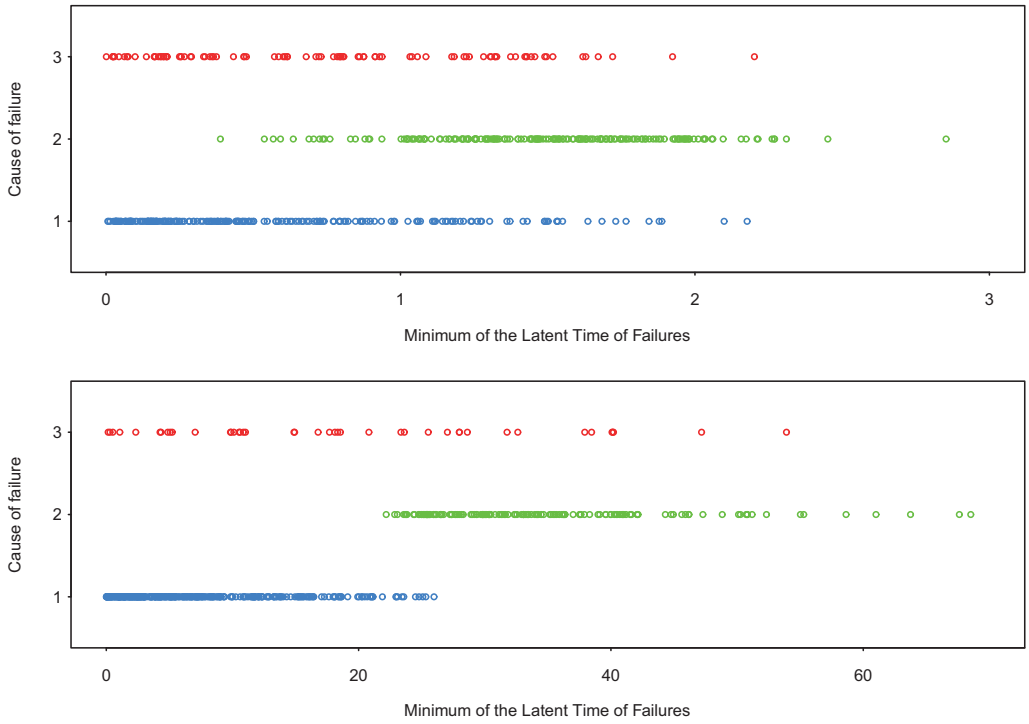


Figure 3 *Dot plot of the minimum between the latent times by cause of failure. Simulation 1 in the upper panel and Simulation 2 in the lower panel. Cause of failure: (1) - 1st cause; (2) - 2nd cause; and (3) - censored value.*

with and without censoring. The copula parameter was often estimated in the border of the parametric space for sample sizes up to 500. The result was worst with censoring, when in many cases the independent copula was selected by the AIC criterion. When the sample size was increased to $1,000$, the AIC criterion seldom selected the independent copula, and the nonparametric and the parametric estimation (by the correct Clayton-Llog-Lnor model) were close to the theoretical hazard function. Even when the wrong copula was fitted, the fit was good, except in the right tail. The reasons for this are the same as in example 1: few observations in the extreme and incorrect tail dependency.

The simulation was also conducted with different copula parameter values. The copula parameter was chosen to have Kendall's τ equal to 0.7, 0.5 and 0.3. When τ is equal to 0.3 or 0.5, the likelihood of the models with independent copula was very close to that of models with dependent copulas, and in general, the AIC criterion selected the independent copula. For $\tau = 0.7$, the AIC criterion almost always selected a dependent copula.

15

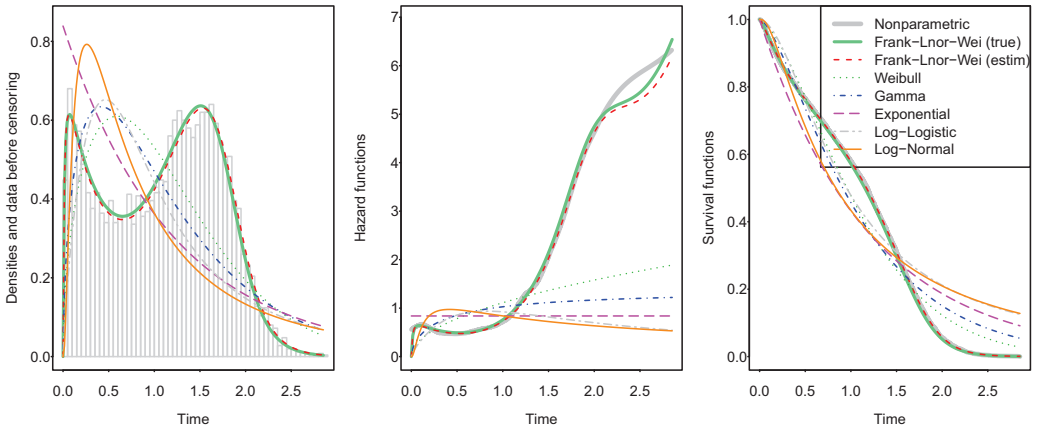


Figure 4 *Simulation 1. Comparison of the estimates density, hazard and survival functions by single risk models and by the polyhazard models of Table 1.*

4.2 Unemployment duration data

The unemployment duration dataset was previously studied by Wichert and Wilke (2008), who described it as "it is a sample of German administrative individual unemployment duration data. It is extracted from the IAB-Employment Sample 1975-2001 (IABS-R01), which contains employment trajectories of about 1.1 million individuals from West-Germany and about $200K$ individuals from East-Germany. It is a 2% random sample of the socially insured workforce." At the time the data were collected, certain rules governed the administration of the two basic benefits related to unemployment: the unemployment benefit and unemployment assistance. The unemployment benefit was granted at the beginning of the individual's unemployment and could last from six to 32 months. The benefit had mechanisms to incentivize the insured individual's return to the job market, for instance, by suspending the benefit of a person who refused a job offer that would pay a salary comparable with that of his or her last job. The unemployment assistance could be granted immediately after the end of the unemployment benefit; it had additional criteria for eligibility, its value was lower than that of the unemployment benefit and it could last indefinitely in time.

The available data consist of the duration of the withdrawals of an individual from one or both of the benefits. Therefore, the date when an individual began and finished his or her withdrawals from the unemployment insurance is the only available measure. The end of the benefit may occur due to several causes, such as emigration, finding another job or starting a business, but this information is not available. Thus, we believe that there are risks

16

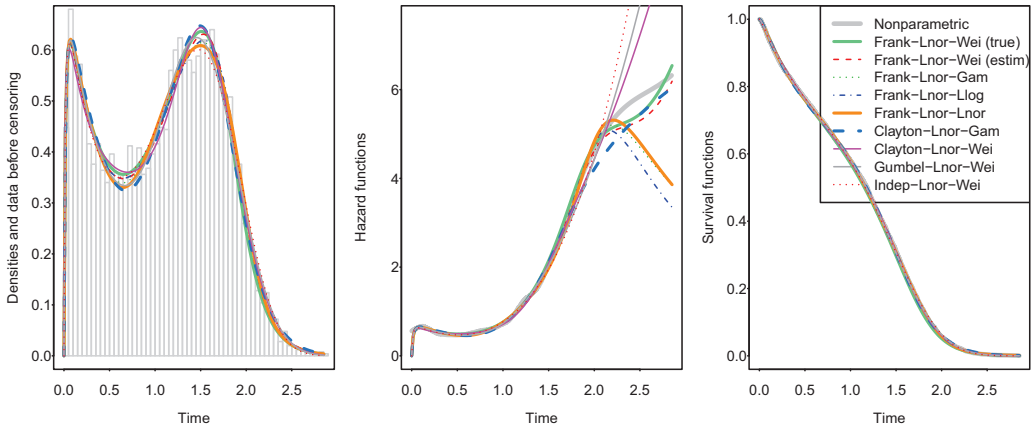


Figure 5 *Simulation 1. Comparison among the best copula model fitted. Density, hazard and survival functions for the polyhazard models of Table 1.*

competing for the end of the unemployment duration of an individual. Only the 8,109 observations of women in the dataset were used. We considered as censored observation cases when the woman was still unemployed by the end of the observation period (the year of 2001) or when she was unemployed when the benefit reached its maximum duration. There are 15.8% censored observations.

Table 3 shows the estimates for each copula for the best AIC polyhazard models fitted to the unemployment data; estimates for the single risk models are also provided. Estimates of the density, hazard and survival functions are presented in Figure 8. The polyhazard models exhibit a good fit to the data, and are clearly superior to the single risk models. The estimated hazard function has a peak at the beginning and a maximum at approximately 1.4 months followed by a subsequent decline. A minimum value is reached at approximately one year and four months, after wich the function increases again. Except for the model with the Frank copula, the estimates show dependence between the latent variables. Independently of the model, the estimates of the density, hazard and survival functions are very close, showing again that the estimation of these functions is robust to the model misspecification.

5 Final remarks

Independent polyhazard models are known to be a flexible tool for the construction of hazard functions. The use of copulas to model the dependence of the latent factors considerably increases this flexibility. With generalized

17

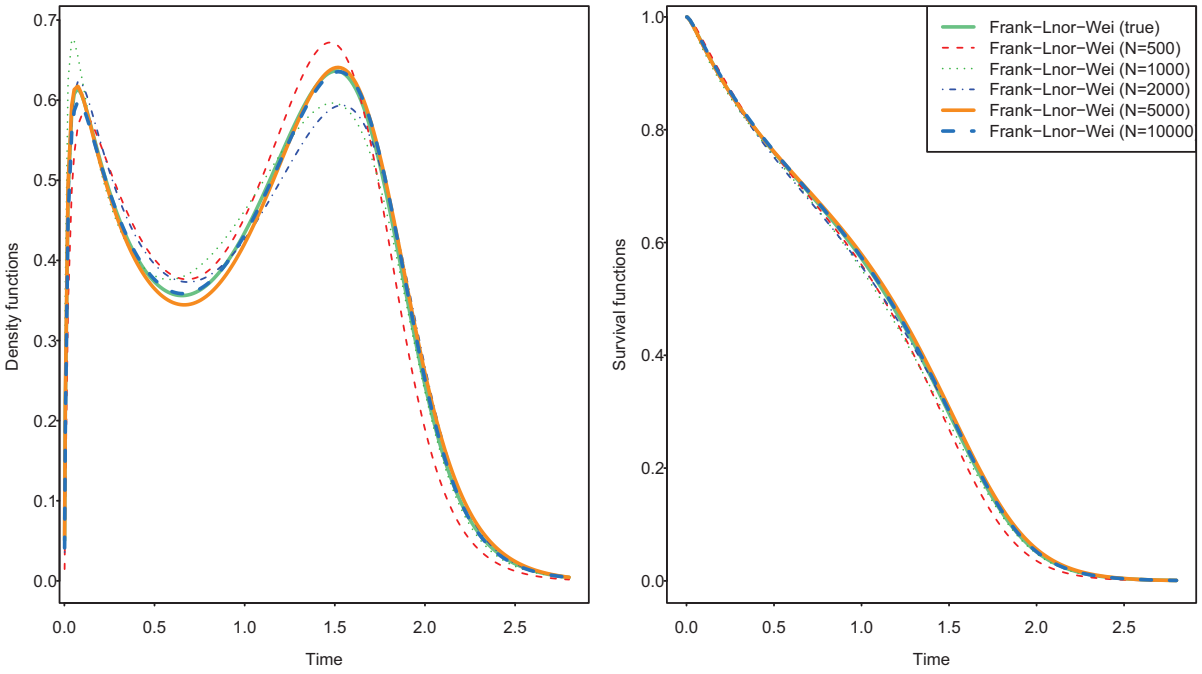


Figure 6 *Simulation 1. Frank-Lnor-Wei model fitted to samples of sizes 500, 1,000, 2,000, 5,000 and 10,000.*

polyhazard models, it is possible to construct a rich family of hazard rate functions with bathtub and multimodal shapes as well as local effects. The proposed model yields a strong fit to simulated data and unemployment duration data representing effects resulting from the presence of competing risks. Although it was not possible to infer the latent times due to the identification issue resulting from the lack of information about the cause of failure, the proposed model conveniently allows for restrictions on dependence (negative, positive or tail dependence), and also allows for the direct examination of the association between covariates and the behavior of the latent times.

Acknowledgments: The authors would like to thank two anonymous referee for carefully readind the paper and for their comments which greatly improved the paper. We also thank Epifisma Laboratory (UNICAMP). This work was partially supported by grants from CNPq, CAPES and FAPESP.

18

Table 2 *Simulation 2. True model: Clayton copula* $\theta = 18$ *(Kendall's* $\tau = 0.90$*) and log-logistic* $(13.0; 1.0)$ *and log-normal* $(3.0; 0.5)$ *marginals, and sample size equal to* $1,000$*. Single risk models and models selected by AIC criterion: best polyhazard models, best marginals configuration for each copula and the copula model for the right marginal configuration*

Model	AIC	τ	θ	Marg. Distrib. 1		Marg. Distrib. 2	
Frank-Llog-Llog	7035.63	0.77 (0.33;0.87)	15.60 (3.24;27.96)	12.95 (11.41;14.49)	0.97 (0.90;1.04)	22.89 (21.00;24.78)	4.65 (3.93;5.37)
Clayton-Llog-Llog	7035.72	0.76 (0.54;0.84)	6.28 (2.38;10.17)	13.03 (11.46;14.60)	0.97 (0.90;1.04)	22.38 (21.03;23.73)	4.55 (3.88;5.22)
Gumbel-Llog-Llog	7035.73	0.83 (-0.02;0.91)	5.87 (0.98;10.75)	12.94 (11.41;14.46)	0.97 (0.90;1.04)	22.88 (20.84;24.92)	4.70 (3.85;5.55)
Clayton-Llog-Lnor	7037.27	0.83 (0.62;0.89)	9.86 (3.21;16.50)	12.92 (11.38;14.45)	0.97 (0.90;1.04)	3.04 (2.98;3.10)	0.44 (0.39;0.49)
Gumbel-Llog-Lnor	7037.69	0.91 (0.62;0.95)	10.98 (2.60;19.35)	12.85 (11.35;14.36)	0.98 (0.91;1.04)	3.04 (2.98;3.11)	0.44 (0.39;0.49)
Frank-Llog-Lnor	7037.75	0.87 (0.52;0.93)	29.61 (6.03;53.19)	12.87 (11.36;14.37)	0.98 (0.91;1.04)	3.05 (2.98;3.11)	0.44 (0.39;0.49)
Indep-Llog-Lnor	7049.22			32.79 (31.39;34.19)	5.85 (5.07;6.62)	2.64 (2.50;2.78)	1.88 (1.76;2.01)
Gamma	7181.87			0.93 (0.86;1.00)	18.80 (16.82;20.77)		
Weibull	7184.91			17.30 (16.14;18.47)	0.98 (0.93;1.03)		
Exponential	7183.37			17.41 (16.29;18.52)			
Log-logistic	7378.90			10.87 (9.93;11.81)	1.27 (1.20;1.33)		
Log-normal	7400.32			2.25 (2.16;2.34)	1.41 (1.35;1.48)		

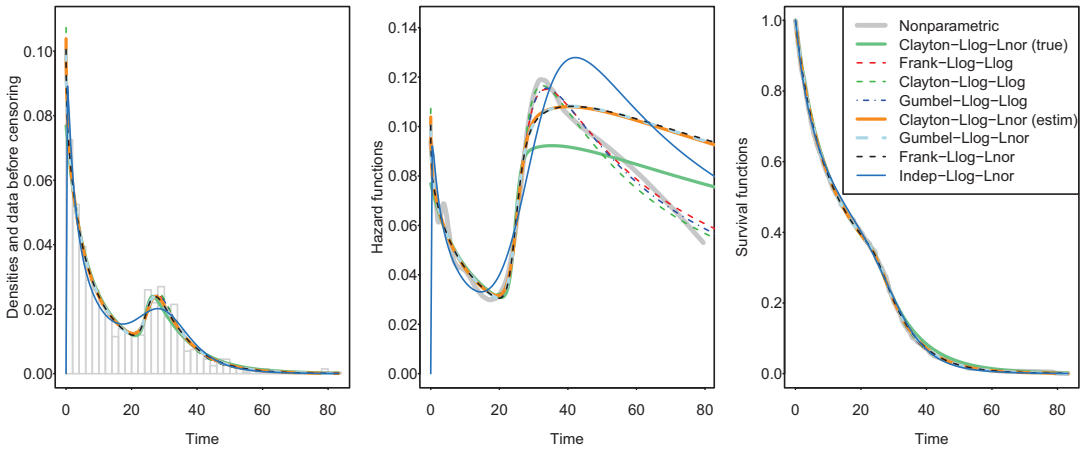


Figure 7 *Simulation 2: Comparison among the best copula model fitted. Density, hazard and survival functions for the polyhazard models of Table 2.*

Table 3 *Summary of the models fitted to the unemployment data. Single risk models and selected polyhazard models. For each copula, only the specification selected by the AIC criterion is presented.*

Model	AIC	τ	θ	Marg. Distrib. 1		Marg. Distrib. 2	
Clayton-Lnor-Gam	20429.48	0.75 (0.68;0.79)	5.90 (4.34;7.45)	0.24 (0.15;0.32)	1.62 (1.56;1.68)	1.45 (1.33;1.57)	1.31 (1.22;1.41)
Gumbel-Lnor-Lnor	20436.03	0.53 (0.00;0.82)	2.14 (1.00;5.44)	0.85 (0.12;1.58)	0.55 (0.35;0.75)	0.13 (0.08;0.17)	1.65 (1.61;1.69)
Frank-Lnor-Lnor	20436.46	-0.05 (-0.28;0.19)	-0.44 (-2.69;1.81)	1.38 (1.11;1.65)	0.50 (0.42;0.57)	0.13 (0.08;0.18)	1.65 (1.61;1.69)
Indep-Lnor-Lnor	20434.62			0.13 (0.08;0.18)	1.65 (1.61;1.70)	1.33 (1.29;1.37)	0.48 (0.45;0.52)
Weibull	20822.76			1.66 (1.62;1.71)	0.92 (0.90;0.93)		
Gamma	20832.89			0.88 (0.86;0.91)	1.95 (1.87;2.03)		
Exponential	20906.22			1.70 (1.66;1.74)			
Log-normal	21170.94			-0.08 (-0.11;-0.05)	1.40 (1.38;1.42)		
Log-logistic	21333.81			0.99 (0.96;1.02)	1.23 (1.20;1.25)		

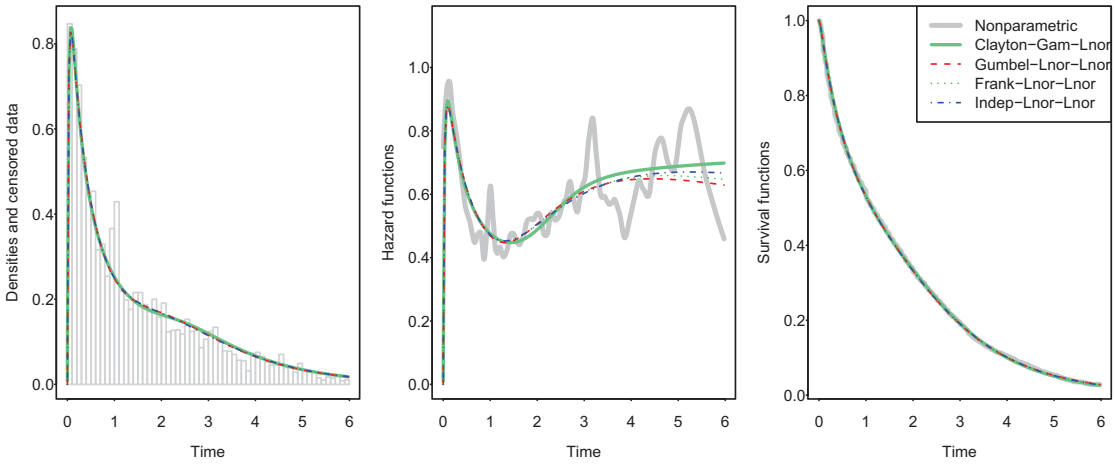


Figure 8 *Density, hazard and survival functions of the models fitted to the women unemployment data. Polyhazard models of Table 3.*

20

References

Basu, S., Basu, A. P. and Mukhopadhyay, C. (1999). Bayesian analysis for masked system failure data using non-identical Weibull Models. *Journal of Statistical Planning and Inference* **78** 255-275. MR1705552

Berger, J. M. and Sun, D. O. (1993). Bayesian analysis for the poly-Weibull distribution. *Journal of the American Statistical Association* **88** 1412-1418. MR1245378

Carriere, J. (1994). Dependent decrement theory. *Transactions of the Society of Actuaries* **46** 45-74.

Cherubini, U., Luciano, E. and Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons. MR2250804

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society B* **34** 187-220. MR0341758

Heckman, J. J. and Honoré, B. E. (1989). The identifiability of the competing risks model. *Biometrika* **76** 325-330. MR1016023

Joe, H. (1997). *Multivariate models and dependence concepts*. Chapman and Hall/CRC. MR1462613

Kaishev, V. K., Dimitrova, D. S. and Haberman, S. (2007). Modelling the joint distribution of competing risks survival times using copula functions. *Insurance Mathematics and Economics* **41** 339-361. MR2364559

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The statistical analysis of failure time data*. Wiley. MR570114

Kuo, L. and Yang, T. M. (2000). Bayesian reliability modeling for masked system lifetime. *Statistics and Probability Letters* **47** 229-241. MR1747483

Louzada-Neto, F. (1999). Polyhazard models for lifetime data. *Biometrics* **55** 1281-1285.

Louzada-Neto, F., Andrade, C. S. and Almeida, F. R. Z. (2004). On the non-identifiability problem arising on the poly-Weibull model. *Communications in Statistics* **33(3)** 541-552. MR2090953

Mazucheli, J., Louzada-Neto, F. and Achcar, J. A. (2001). Bayesian inference for polyhazard models in the presence of covariates. *Computational Statistics & Data Analysis* **38** 1-14. MR1869477

Nadarajah, S., Cordeiro, G. M. and Ortega, E. M. M. (2011). General results for the beta modified Weibull distribution. *Journal of Statistical Computation and Simulation* **81** 1211-1232.

Nelsen, R. B. (2006). *An introduction to copulas*, 2 ed. Springer. MR2197664

Pham, H. and Lai, C. D. (2007). On recent generalizations of the Weibull distribution. *IEEE Transactions on Reliability* **56** 454-458.

Trivedi, P. K. and Zimmer, D. M. (2005). Copula modelling: An introduction for practitioners. *Foundations and Trends in Econometrics* **1** 1-111.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences USA* **72** 20-22. MR0356425

Wichert, L. and Wilke, R. A. (2008). Simple non-parametric estimators for unemployment duration analysis. *Journal of the Royal Statistical Society - Series C* **1** 117-126. MR2412670

Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika* **82(1)** 127-138. MR1332844

Apêndice B

Fitting Distributions with the Polyhazard Model with Dependence

Fitting Distributions with the Polyhazard Model with Dependence

Rodrigo Tsai e Luiz Koodi Hotta

A aparecer na revista *Communications in Statistics - Theory and Methods*
http://www.tandfonline.com/toc/lsta20/current/

Fitting Distributions with the Polyhazard Model with Dependence

Rodrigo Tsai[a,b] **and Luiz K. Hotta**[a]

[a] *State University of Campinas - UNICAMP*

[b] *Superior Court of Justice - STJ/Brazil*

Abstract

The polyhazard model with dependent causes, first introduced to fit lifetime data, generalized the traditional polyhazard model by allowing the latent causes of failure to be dependent by using copula functions. When modeling lifetime data, marginal distributions are supported on the positive reals. Dropping this restriction, the method generates a rich family of univariate distributions with asymmetries and multiple modes. We show that this new family of distributions is able to approximate other distributions proposed in the literature, such as the generalized beta-generated distributions. These distributions are fitted to three real data sets.

Keywords: Polyhazard model, Generalized beta generated models, Copula, Family of distributions

Mathematics Subject Classification: 62N99, 62H99, 62E99

1 Introduction

The polyhazard model with dependent causes (PHD) was first introduced by Tsai and Hotta (2011) to fit lifetime data. The model generalized the traditional polyhazard model by allowing the latent causes of failures to be dependent. The dependence is modeled through a copula function. The main feature of this model is its ability to represent hazard functions with unusual shapes, such as bathtub or multimodal curves, while also modeling local effects. When modeling lifetime data, marginal distributions are supported on the positive reals. Dropping this restriction, the method can be used to generate a rich family of univariate distributions with asymmetries and multiple modes. Many methods have been proposed in the literature to generate univariate distributions. One rich family is the generalized beta-generated (GBG) distributions (Alexander et al., 2012; Cordeiro et al., 2012). The method proposed here is able to approximate most of the distributions from the GBG family. The remainder of this paper is organized as follows. In Sec. 2, we will briefly introduce the PHD model. The comparison between the new family and the GBG family will be presented in Sec. 3. In Sec. 4, the PHD model will be applied to data previously analyzed in the literature using GBG models. Finally, some concluding remarks will be provided in Sec. 5.

2 The polyhazard model with dependent causes

We first introduce the PHD model as proposed for lifetime applications. Suppose that the survival of one subject is subjected to $k \geq 2$ competing latent causes of failure. Let the lifetime related to the jth latent cause of the ith unit of

observation, X_{ij}, have density $f_j(\cdot;\Gamma_j)$, which is considered as known, except for the unknown set of parameters Γ_j. Denote the survival and hazard functions by $S_j(\cdot;\Gamma_j)$ and $\lambda_j(\cdot;\Gamma_j)$. In this set up only $X_i = \min\{X_{ij}, j = 1, \cdots, k\}$ is observed for each unit of observation. Thus, assuming the risks are independent, i.e., the failure times X_{ij}, $j = 1, \cdots, k$ are independent, the overall survival function of X_i, denoted by $S(t;\Upsilon)$, where $\Upsilon = (\Gamma_1, \cdots, \Gamma_k)$, is given for any $i = 1, \cdots, n$ by the product of marginal survival functions:

$$S(t;\Upsilon) = P_\Upsilon[X_i > t] = P_\Upsilon[X_{i1} > t, \cdots, X_{ik} > t] = \prod_{j=1}^{k} S_j(t;\Gamma_j), \tag{1}$$

and the hazard function of X_i, $\lambda(t;\Upsilon)$, is given by the sum of the marginal hazards:

$$\lambda(t;\Upsilon) = \sum_{j=1}^{k} \lambda_j(t;\Gamma_j). \tag{2}$$

In this paper we model the failure time X_i by k competing risks, allowing for dependence between the risks. Henceforth, we use the notation for $k = 2$ for simplicity, but the notation can be easily generalized to $k > 2$. Denoting by $H(\cdot,\cdot;\Upsilon)$ the joint distribution function and by $\bar{H}(\cdot,\cdot;\Upsilon)$ the joint survival function of the latent variables X_{i1} and X_{i2}, we can write the survival function of X_i as

$$S(t;\Upsilon) = P_\Upsilon[X_{i1} > t, X_{i2} > t] = \bar{H}(t,t;\Upsilon). \tag{3}$$

To model the joint survival function $\bar{H}$, allowing for dependence between the latent variables, we propose the use of copula functions. A k-dimensional copula function may be defined as a cumulative distribution function whose marginal distributions are uniform over $[0,1]$ and whose support is the $[0,1]^k$ hypercube. Copula functions have been extensively studied in the multivariate modeling literature, especially when the use of the multivariate normal distribution is questionable. An important feature of the copula approach is the possibility of modeling the dependence and the marginal behavior of the related variables separately, thus making the copula a very convenient alternative in the case of multivariate modeling. For further details on copulas look at Joe (1997), Cherubini et al. (2004), Trivedi and Zimmer (2005) and Nelsen (2006).

Let $F_1(\cdot;\Gamma_1)$ and $F_2(\cdot;\Gamma_2)$ be the distribution functions of X_{i1} and X_{i2}, respectively. It follows from Sklar's theorem that there is always a copula function C^* such that we can write $H(t_1,t_2;\Upsilon) = C^*(F_1(t_1;\Gamma_1), F_2(t_2;\Gamma_2))$, and that C^* is unique if the marginal distributions F_1 and F_2 are continuous. C^* is then called a copula function because it couples the marginal distributions F_1 and F_2 to their joint distribution H. It is possible to represent the joint survival function directly by $\bar{H}(t_1,t_2;\Upsilon) = P[X_{i1} > t_1, X_{i2} > t_2;\Upsilon] = \tilde{C}(S_1(t_1;\Gamma_1), S_2(t_2;\Gamma_2))$, where $\tilde{C}(u,v) = u + v - 1 + C^*(1-u, 1-v)$ is also a copula. On the other hand, for C any copula, $C(S_1(t_1;\Gamma_1), S_2(t_2;\Gamma_2))$ is a survival distribution function. Therefore we can also model the survival function S directly by a copula function C as in Kaishev et al. (2007). This is also the approach adopted here, since it is generally easier to work analytically with this representation. Then, for the survival function of the polyhazard model with dependence given by a copula function C with dependence parameter θ and $\Upsilon = (\theta, \Gamma_1, \Gamma_2)$, we can write

$$S(t;\Upsilon) = \bar{H}(t,t;\Upsilon) = C_\theta(S_1(t;\Gamma_1), S_2(t;\Gamma_2)), \tag{4}$$

where S_1 and S_2 are, in this paper and in almost all practical applications, continuous marginal survival functions. The copula C in (4) is called the survival copula: in this paper, we refer to it as the copula function. From the survival function (4), it follows that the probability density and hazard functions for the polyhazard model with dependence

are obtained in the usual fashion, i.e.,

$$f(t;\Upsilon) = -\frac{d}{dt}S(t;\Upsilon) \quad \text{and} \quad h(t;\Upsilon) = \frac{f(t;\Upsilon)}{S(t;\Upsilon)}. \tag{5}$$

Dropping the restriction that the latent variables are positive, we can no longer talk about lifetime distributions, and the concepts of survival function and hazard function are no longer valid, either. On the other hand, we have a new method of generating a rich family of distributions. In the next section, we will compare this family with another family of distributions found in the literature.

3 Theoretical comparison

In this section, we will compare two families of distributions, i.e., check whether distributions generated by one family can be approximated by distributions from the other family. We measure their closeness by comparing their density functions. In order to also compare the hazard functions, in this section we will only work with non-negative distributions.

A GBG distribution is defined by the generalized beta distribution of the first kind and by a parent distribution $F(x;\Theta)$ and its density $f(x;\Theta)$, where Θ is a vector of parameters (Alexander et al., 2012).

The probability density function of the generalized beta distribution of the first kind, with positive parameters a, b and c, is

$$f_{GB}(u;a,b,c) = cB(a,b)^{-1}u^{ac-1}(1-u^c)^{b-1}, \quad 0 < u < 1,$$

where $B(a,b)$ is the beta function with parameteres a and b. The density of the GBG distribution is

$$f_{GBG}(x;\Theta,a,b,c) = cB(a,b)^{-1}f(x;\Theta)F(x;\Theta)^{ac-1}[1-F(x;\Theta)^c]^{b-1}, \ x \in I,$$

where I is the support of $f(x;\Theta)$. For each parent distribution, we have a GBG distribution. We will consider here the following GBG distributions: Weibull (GBW) (GBG distribution with Weibull parent distribution), normal (GBN), log-normal (GBLN), gamma (GBGam), Gumbel (GBGum), two degrees of freedom skewed-t (GBTSk), and scaled Laplace (GBSLa). We also consider the beta-modified Weibull distribution (BMW) (Nadarajah et al., 2011), which is defined by the distribution function

$$F(x;a,b,\alpha,\beta,\lambda) = \frac{1}{B(a,b)}\int_0^{1-\exp\{\alpha x^\beta \exp(\lambda x)\}} w^{a-1}(1-w)^{b-1}dw$$

and the beta generalized gamma distribution (BGGam) introduced by Cordeiro et al. (2012), given by

$$F(x;a,b,c,k,\lambda) = \frac{1}{B(a,b)}\int_0^{\gamma(k,(\lambda x)^c)/\Gamma(k)} w^{a-1}(1-w)^{b-1}dw,$$

where $\gamma(\beta,x) = \int_0^x w^{\beta-1}e^{-w}dw$ is the incomplete gamma function. Henceforth we refer to the set of distributions given by the presented beta integration method, such as GBG, BMW, and BGGam, as beta generated (BG) models.

In the next section we introduce the method, and in Sec. 3.2, the true distribution is from the PHD family, and in Sec. 3.3, the true distribution will belong to the BG family. In the first case, we consider two distributions presented in Tsai and Hotta (2011) that introduced the PHD model, and six BG distributions from Alexander et al. (2011, 2012), Nadarajah et al. (2011), and Cordeiro et al. (2012). Among the many examples in these papers, we selected the functions exhibiting multimodality and/or local effects in their density and/or hazard functions.

3.1 The method

Let $f_{ph}(.)$ and $h_{ph}(.)$ be the density and hazard functions, respectively, selected from the PHD family. We will first look for distributions from the BG family which have a density close to the PHD density function, and then repeat the same exercise for the hazard function. Let $f(.;\Upsilon)$ and $h(.;\Upsilon)$ be the density and hazard functions for the BG family with parameters Υ. Define the objective functions

$$f_{Obj1}(x,\Upsilon) = \sum_{i=1}^{N}[f_{ph}(x_i) - f(x_i;\Upsilon)]^2 \tag{6}$$

and

$$f_{Obj2}(x,\Upsilon) = \sum_{i=1}^{N}[h_{ph}(x_i) - h(x_i;\Upsilon)]^2. \tag{7}$$

For each problem, either density or hazard function analysis, we search for a value of Υ that minimizes the objective function in Υ. Let $x = (x_1, \cdots, x_N)$ and $y = (y_1, \cdots, y_N)$ be a grid of N points where the functions are to be compared. The density and the hazard functions for the PHD model f_{ph} and h_{ph} were taken from $t = 0$ to the 99.9% percentile of the distribution. The use of an equally spaced grid turned out to be inappropriate, and we decided to choose the grid empirically, but respecting a criterion in relation to both axes, i.e., with $x_i - x_{i-1} < dx$ and $|f_{ph}(x_i) - f_{ph}(x_{i-1})| < df$ with $dx > 0$ and $df > 0$ for the density function and $y_i - y_{i-1} < dy$ and $|h_{ph}(y_i) - h_{ph}(y_{i-1})| < dh$ with $dy > 0$ and $dh > 0$ for the hazard function. Here, dx, df, dy, and dh depended on the functions studied. The optimization procedures were performed in R software (R Core Team, 2012), using the "optim" function with the "Nelder-Mead" method. Also, in each optimization we used $10,000$ points uniformly selected in the parametric space of the generalized models as initial values in the optimization. The parametric space considered for each BG model was the hyper-rectangle given by the Cartesian product of intervals of the model parameters.

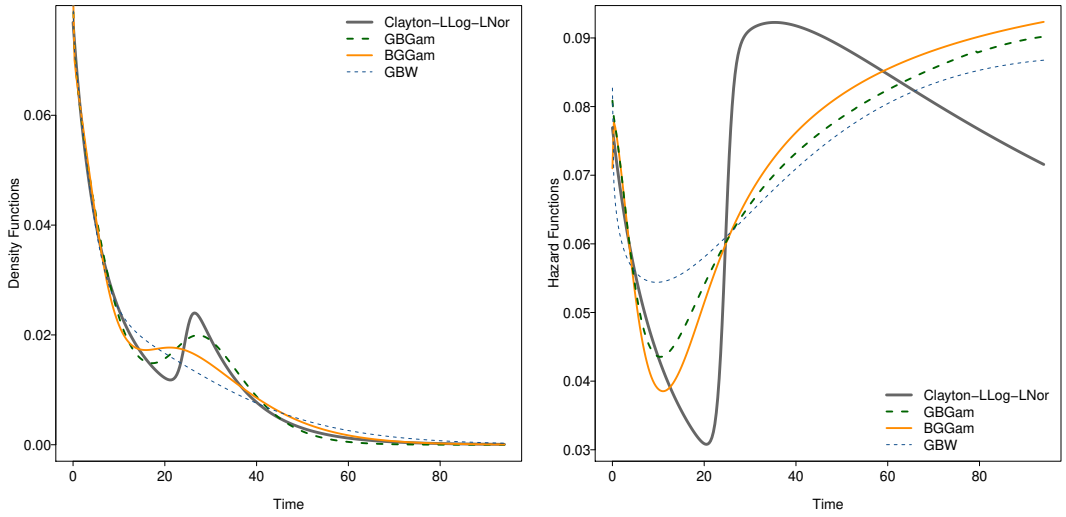
For the generalized beta, the parameters a and b were taken in the range $[0; 30]$ and the shape parameter c in the interval $[0; 10]$. The parametric intervals for the GBN distribution were $[-50; 50]$ for the location parameter, and $[0; 20]$ for the scale parameter; for the GBLN distribution, they were $[-50; 30]$ for the location parameter, and $[0; 20]$ for the scale parameter; for GBGam, they were $[0; 20]$ for both parameters; for GBGum, they were $[0; 20]$ for both parameters; for GBTSk, it was also $[0; 20]$ for the scale factor parameter; for GBSLa, it was $[0; 20]$ for the scale parameter; for GBW and BMW, they were $[0; 10]$ for both parameters.

In a few cases, we generated a very large sample $(10,000)$ from some BG distributions and fitted some PHD distributions to the data. The maximum likelihood estimates were very close to the values found in the above method. This supports the assertion that this method is able to find a density from one family that is a good approximation to a density from another family.

3.2 Distributions generated by the PH family

In this section, the true distributions belong to the PHD family and we look for a distribution from the BG family to approximate them. In the first example, we have a Clayton copula with log-logistic and log-normal latent distributions (Clayton-LLog-Lnor). Figure 1 shows the results of the best three adjustments for the density and hazard function from the BG distributions. We do the same with the PHD with the Frank copula and log-normal and Weibull latent distributions (Frank-Lnor-Wei). The results are shown in Figure 2. The PHD models were taken from Tsai and Hotta

(2011). The analyses of both examples show that the BG family of distributions is not able to approximate any of these distributions, neither the density nor the hazard function.



(a) (b)

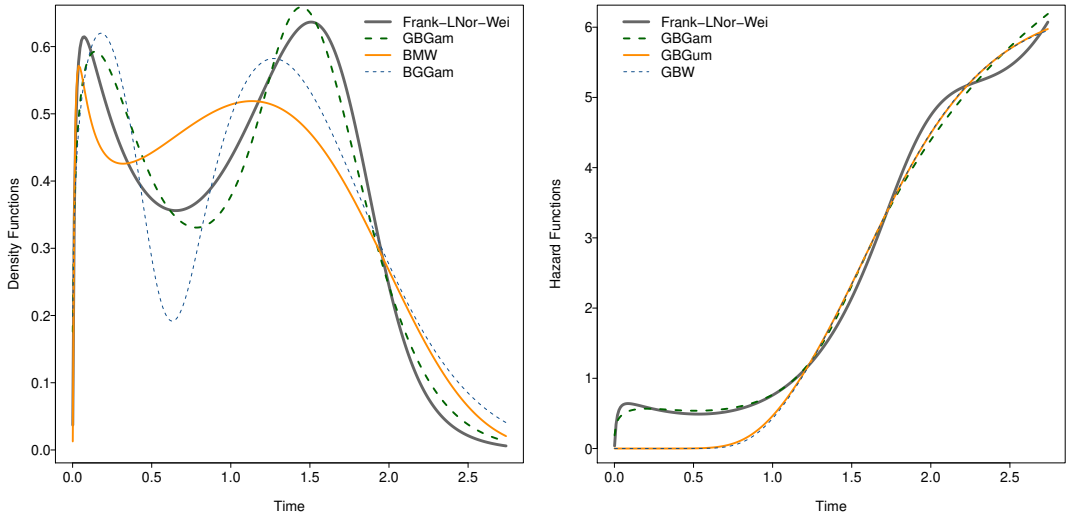
Figure 1: Best adjustment of a PHD Clayton-LLog-Lnor density and hazard functions by BG family density and hazard functions.

3.3 Distributions from the BG family

In this section, we will study whether the PHD family of distributions is able to approximate distributions from the BG family. In the PHD family, we only consider the Frank, Clayton, Gumbel, and the symmetrized Joe-Clayton (SJC) copulas, and for the marginals, the normal (Nor), Weibull (Wei), log-normal (Lnor), log-logistic (llog), Gamma (Gam), and the extreme value (Eva) distributions. We must take into account that when the observations can only take positive values, we should only use distributions supported on positive reals.

For the hazard function representation task, we first selected the following BG distributions: BMW with parameters $a = 1.0$, $b = 0.7$, $\alpha = 1.0$, $\beta = 0.3$, and $\mu = 10.0$ (Figure 2 in Nadarajah et al., 2011); GBW with parameters $a = 2.0$, $b = 2.5$, $c = 1.2$, $\beta = 0.65$, and $\mu = 0.5$ (Figure 7 in Alexander et al., 2011); and GBW with parameters $a = 1.0$, $b = 1.5$, $c = 0.05$, $\beta = 3.0$, and $\mu = 2.0$ (Figure 7 in Alexander et al., 2011). For these cases, several models from the PHD family approximate well the true hazard functions. Figure 3 presents the three best adjustments of each of the hazard functions.

The second set of distributions for the density function representation are from the distributions: BGGam with parameters $a = 0.1305$, $b = 0.0185$, $c = 6.0008$, $k = 5.9155$, and $\mu = 0.8229$, which models the fiber data in Cordeiro et al. (2012); GBW with parameters $a = 0.07$, $b = 0.1$, $c = 10.0$, $\beta = 9.0$, and $\mu = 1.25$ (Figure 6 in Alexander et al., 2012); and GBN with parameters $a = 0.02$, $b = 0.02$, $c = 0.5$, $\mu = 0.0$ and $\sigma = 1.0$ (Figure 5 in Alexander et al., 2012). Figure 4 presents the three best adjustments of each of the density functions. For the density given in Figure



(a) (b)

Figure 2: Best adjustment of a PHD Frank-Lnor-Wei density and hazard functions by BG family density and hazard functions.

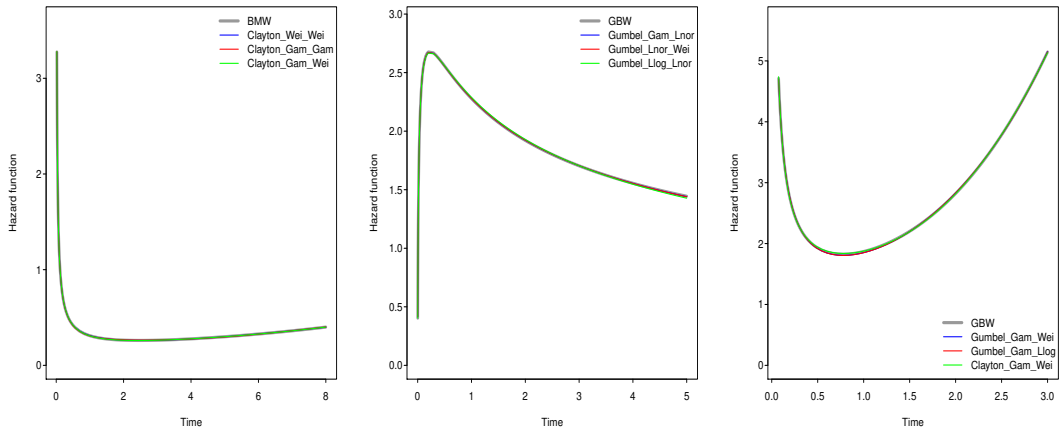
5 (a) the adjustment was good, but for the densities (b) and (c), especially for the latest one, the adjustments were not so good.

4 Empirical Applications

PHD densities will be fitted here to three data sets from the literature. In all the original applications, the authors fitted densities from the BG family. When it makes sense, we will also show the estimated hazard and survival functions. The data sets considered are the voltage data that was previously studied by Meeker and Escobar (1998), the skew normal data, previously considered in Sec. 6.3 of Azzalini (1999) and in Sartori (2006), and the fiber data analyzed by Smith and Naylor (1987).

Voltage data

This data set was previously studied by Meeker and Escobar (1998, p. 383). Alexander et al. (2012) also analyzed the data, fitting a GBW distribution, and described the data as follows: "These data represent the times of failure and running times for a sample of devices from a field-tracking study of a larger system. At a certain point in time 30 units were installed in normal service conditions. Two causes of failure were observed for each unit that failed: the failure caused by an accumulation of randomly occurring damage from power-line voltage spikes during electric storms and failure caused by normal product wear." We fitted the densities taking into account that there are eight censored observations at value 300. Alexander et al. (2012) did not considered that these observations



(a) (b) (c)

Figure 3: Best adjustment of Beta-Modified Weibull hazard functions by PHD family hazard functions. (a) BMW ($a = 1.0$; $b = 0.7$; $\alpha = 1.0$; $\beta = 0.3$; $\mu = 10.0$); (b) GBW ($a = 2.0$; $b = 2.5$; $c = 1.2$; $\beta = 0.65$; $\mu = 0.5$); (c) GBW ($a = 1.0$; $b = 1.5$; $c = 0.05$; $\beta = 3.0$; $\mu = 2.0$).

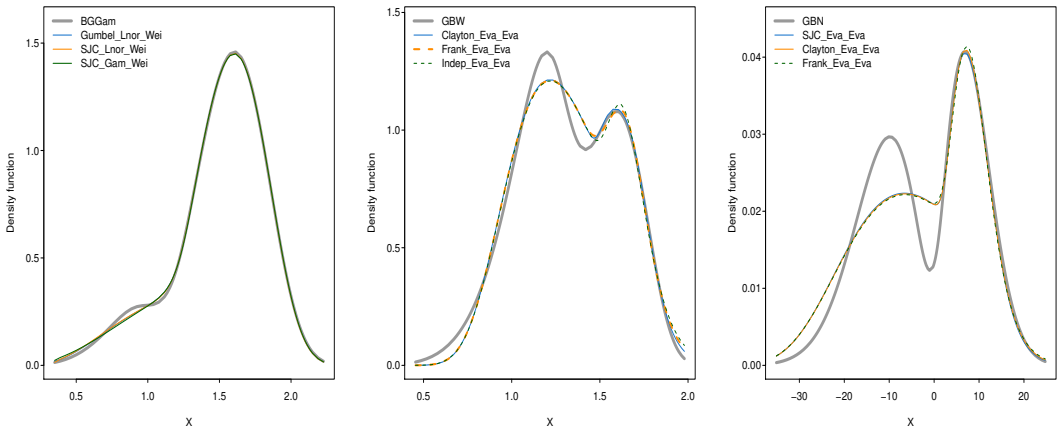
are censored values. Figure 5 (a) shows the adjusted density functions for the three PHD models with best AIC. Although Alexander et al. (2012) did not consider the censoring, their estimate is presented in the figure. For all the PHD models, we constrained the copula parameter to have Kendall's τ and tail dependence coefficient less than or equal to 0.70. This smoothing restriction was done because we have only 30 observations and this makes the latent components of the model highly dependent on a few observations. This is not expected to happen with larger samples.

Skew normal data

This data set is a sample from the skew normal distribution with parameters $(\mu, \sigma, \lambda) = (0, 1, 5)$, where λ stands for the shape parameter, to fit the PHD model. This sample was previously considered in Azzalini (1999) and in Sartori (2006), and analyzed with the GBG model in Alexander et al. (2012). Azzalini (1999) had to fix the shape parameter in the estimation of the skew normal distribution because the maximum likelihood method leads to infinity. The sample size was not enough to estimate the parameter. The data is also available at http://azzalini.stat.unipd.it/SN/index.html. Figure 5 (b) shows the fit of three PHD densities and by a GBN density to the data. The PHD densities were selected by the AIC criterion and we constrained the copula parameter to have Kendall's τ less than or equal to 0.70.

Fiber data

The fiber data comes from Smith and Naylor (1987). According to them, "The samples are experimental data of the strength of glass fibers of two lengths, 1.5 cm, and 15 cm, from the National Physical Laboratory in England. Preliminary inspection of the data reveals possible outliers in the lower end of the sample, the smallest observation in



(a) (b) (c)

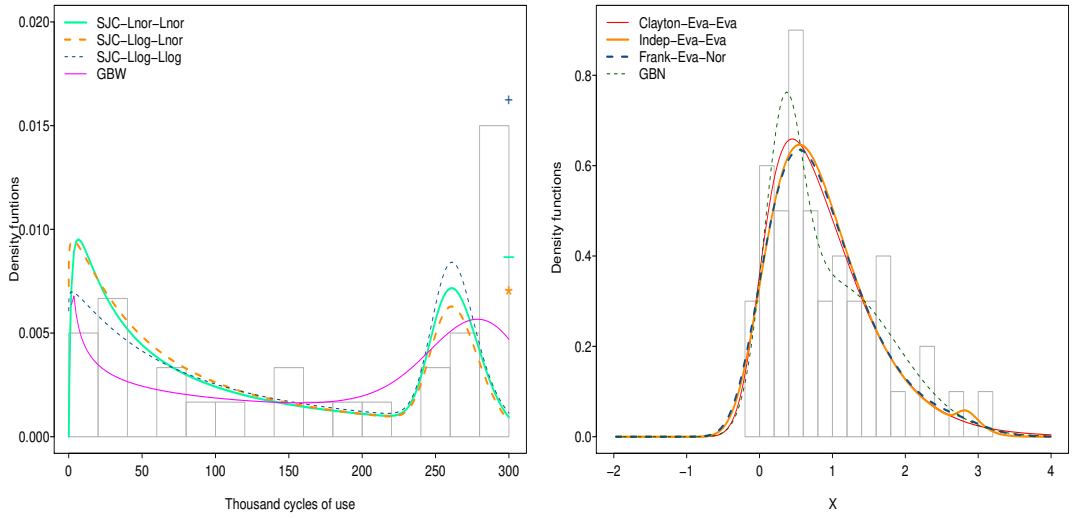
Figure 4: Best adjustment of BG density functions by PHD family density functions. (a) BGGam ($a = 0.1305$, $b = 0.0185$, $c = 6.0008$, $k = 5.9155$, $\mu = 0.8229$); (b) GBW ($a = 0.07$, $b = 0.1$, $c = 10.0$, $\beta = 9.0$, $\mu = 1.25$); (c) GBN ($a = 0.02$, $b = 0.02$, $c = 0.5$, $\mu = 0.0$, $\sigma = 1.0$).

Sample 1 and the smallest two in Sample 2." We use the data from sample 1 with 63 observations (unknown units). Smith and Naylor (1987) fitted a three-parameter Weibull distribution and we fitted a PHD model with the Frank copula with the restriction of Kendall's $\tau \leq 0.70$ with different marginal distributions. The BGGam distribution was fitted with parameters $(a, b, c, k, \mu) = (0.1305, 0.0185, 6.0008, 5.9155, 0.8229)$. The results are shown in Figure 6 with the fitted BGGam model in Cordeiro et al. (2012). In all three examples, the adjustment by PHD densities was very good.

5 Final remarks

In many practical situations, the distribution of a random variable cannot be modeled by the usual densities. Acknowledging this fact, there is a large number of authors proposing new families of distributions. Tsai and Hotta (2011) proposed the PHD model to fit lifetime data. We showed that when we drop the constraint that the latent variables are positive, we are able to generate a rich family of distributions. In particular, we showed that the PHD model can approximate densities from the BG family of distributions. The PHD family of densities fitted well to three data sets found in the literature. The richness of the PHD family could be increased using two-parameter copula functions, by increasing the number of latent variables, or by using distributions from the BG family as latent distributions.

Acknowledgments: This work was partially supported by grants from CNPq, CAPES and FAPESP. We also thank Epifisma Laboratory (UNICAMP).



(a) (b)

Figure 5: (a) Estimated density functions for the voltage data by PHD densities and by the GBW model fitted in Alexander et al. (2012). There are eight censored observations at voltage value equal to 300. The graph also shows the survival function estimated at this value with dash, asterisk and plus signs for the SJC-Lnor-Lnor, SJC-Llog-Lnor and SJC-Llog-Llog models, respectively. (b) Estimated density functions for the skew normal sample data by PHD densities and by the GBN model fitted in Alexander et al. (2012). In both cases, the best three densities of the PHD family selected by AIC criterion are shown.

References

Alexander, C., Cordeiro, G. M., Ortega, E. M. M., and Sarabia, J. M. (2011). Generalized beta-generated distributions. ICMA Centre Discussion Papers in Finance DP2011-05, *University of Reading, UK*.

Alexander, C., Cordeiro, G. M., Ortega, E. M. M., and Sarabia, J. M. (2012). Generalized beta-generated distributions. *Comput. Statist. Data Anal.* 56:1880–1897.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew-normal distribution. *J. Roy. Statist. Soc. Ser. B* 61:579–602. MR:1707862.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula Methods in Finance*. Chichester: John Wiley & Sons. MR:2250804.

Cordeiro, G. M., Castelares, F., Montenegro, L. C., and Castro, M. (2012). The beta generalized gamma distribution. *Statistics* 84:1–13. DOI:10.1080/02331888.2012.658397.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman and Hall/CRC. MR:1462613.

Kaishev, V. K., Dimitrova, D. S., and Haberman, S. (2007). Modelling the joint distribution of competing risks survival times using copula functions. *Insur. Math. Econ.* 41:339–361. MR:2364559.

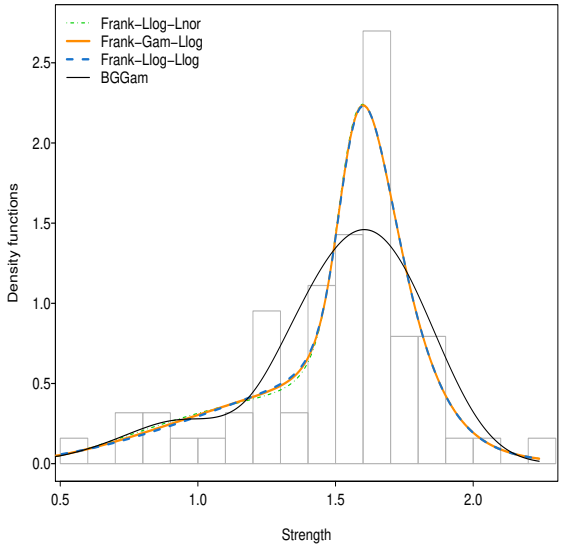


Figure 6: Best adjustment by a density from the PHD density family with Frank copula and adjustment by BGGam density to the fiber data.

Meeker, W. Q. and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. New York: John Wiley & Sons.

Nadarajah, S., Cordeiro, G. M., and Ortega, E. M. M. (2011). General results for the beta modified Weibull distribution. *J. Stat. Comput. Simul.* 81:1211–1232. MR:2843455.

Nelsen, R. B. (2006). *An Introduction to Copulas*. New York: Springer. MR:2197664.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN:3-900051-07-0, URL:http://www.R-project.org/.

Sartori, N. N. (2006). Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *J. Statist. Plann. Inference* 136:4259–4275. MR:2323415.

Smith, R. L. and Naylor, J. C. (1987). A comparison of maximum likelihood and bayesian estimators for the three-parameter Weibull distribution. *J. Roy. Statist. Soc. Ser. C Appl. Stat.* 36:358–369. MR:0918854.

Trivedi, P. K. and Zimmer, D. M. (2005). Copula modelling: An introduction for practitioners. *Found. Trends Econom.* 1:1–111.

Tsai, R. and Hotta, L. K. (2011). Polyhazard models with dependent causes. *Braz. J. Probab. Stat.*, to apear.

Referências Bibliográficas

Akaike (1973) H. Akaike. Information theory and an extension of the maximum likelihood principle. Em *2nd Int. Symp. on Information Theory (Edited by B. N. Petrov and F. Csaki)*, páginas 267–281. Akademiai Kiado. Citado na pág. 45

Alexander *et al.* **(2011)** C. Alexander, G. M. Cordeiro, E. M. M. Ortega e J. M. Sarabia. Generalized beta-generated distributions. *ICMA Centre Discussion Papers in Finance DP2011-05*, 05. Citado na pág. 26, 27, 65

Alexander *et al.* **(2012)** C. Alexander, G. M. Cordeiro, E. M. M. Ortega e J. M. Sarabia. Generalized beta-generated distributions. *Computation Statistics & Data Analysis*, 56:1880–1897. Citado na pág. 26, 27

Azzalini (1986) A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 13:271–275. Citado na pág. 66

Azzalini e Bowman (1990) A. Azzalini e A. W. Bowman. A look at some data on the old faithful geyser. *Journal of the Royal Statistical Society, Series C*, 39:357–365. Citado na pág. 50

Basu *et al.* **(1999)** S. Basu, A. P. Basu e C. Mukhopadhyay. Bayesian analysis for masked system failure data using non-identical weibull models. *Journal of Statistical Planning and Inference*, 78: 255–275. Citado na pág. 19

Bechtel *et al.* **(1993)** Y. C. Bechtel, C. Bonaiti-Pellie, N. Poisson, J. Magnette e P. R. Bechtel. A population and family study n-acetyltransferase using caffeine urinary metabolites. *Clinical Pharmacology and Therapeutics*, 54:134–141. Citado na pág. 50

Berger e Sun (1993) J. M. Berger e D. O. Sun. Bayesian analysis for the poly-weibull distribution. *Journal of the American Statistical Association*, 88:1412–1418. Citado na pág. 1, 19

Berkson e Gage (1952) J. Berkson e R. P. Gage. Survival curves for cancer patients following treatment. *Journal of the American Statistical Association*, 47:501–515. Citado na pág. 15

Boag (1949) J. W. Boag. Maximum likelihood estimates of the proportional of patients cured by cancer therapy. *Journal of the Royal Statistical Society*, 11:15–53. Citado na pág. 15

Bucar *et al.* **(2004)** T. Bucar, M. Nagode e M. Fajdiga. Reliability approximation using finite weibull mixture distributions. *Reliability engineering and system safety*, 84:241–251. Citado na pág. 29

Carriere (1994) J. Carriere. Dependent decrement theory. *Transactions of the Society of Actuaries*, XLVI:45–65. Citado na pág. 25

Cherubini *et al.* **(2004)** U. Cherubini, E. Luciano e W. Vecchiato. *Copula Methods in Finance*. Chichester: John Wiley & Sons. Citado na pág. 3, 20

Clayton (1978) D. G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease. *Communications in Statistics - Theory and Methods*, 29:193–210. Citado na pág. 20

Cordeiro *et al.* **(2012)** G. M. Cordeiro, F. Castelares, L. C. Montenegro e M. De Castro. The beta generalized gamma distribution. *Statistics*, 84:1–13. Citado na pág. 26, 65

Cox (1972) D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34:187–220. Citado na pág. 12, 25

Crawford (1994) S. L. Crawford. An application of the laplace method to finite mixture distributions. *Journal of the American Statistical Association*, 89:259–267. Citado na pág. 48

Crawford *et al.* **(1992)** S. L. Crawford, M. H. DeGroot, J. B. Kadane e M. J. Small. Modeling lake-chemistry distributions: Approximate bayesian methods for estimating a finite-mixture model. *Technometrics*, 34:441–453. Citado na pág. xviii, 48, 49, 62

Georges *et al.* **(2001)** P. Georges, A. G. Lammy, E. Nicolas, G. Quibel e T. Roncalli. Multivariate survival modelling: a unified approach with copulas. *Groupe de Recherch Operationelle, Crédit Lyonnaise*. Citado na pág. 20

Hardle (1990) W. Hardle. *Smoothing Techniques with Implementation in S*. New York: Springer-Verlag. Citado na pág. 50

Heckman e Honoré (1989) J. J. Heckman e B. E. Honoré. The identifiability of the competing risks model. *Biometrika*, 76:325–330. Citado na pág. 25

Henze (1985) N. Henze. A probabilistic representation of the skew normal distribution. *Scandinavian Journal of Statistics*, 12:171–178. Citado na pág. 66

Hosmer *et al.* **(2008)** D. W. Hosmer, S. Lemeshow e S. May. *Applied Survival analysis: Regression Modelling of Time to Event Data*. John Wiley & Sons, 2 ed. Citado na pág. 6

Jewel (1982) N. P. Jewel. Mixtures of exponential distributions. *Annals of Statistics*, 10:479–484. Citado na pág. 29

Jiang e Murthy (1998) R. Jiang e D. N. P. Murthy. Mixture of weibull distributions - parametric characterization of failure rate function. *Applied Stochastic Models and Data Analysis*, 14:47–65. Citado na pág. 29

Joe (1997) H. Joe. *Multivariate Models and Dependence Concepts*. Chapman and Hall/CRC. Citado na pág. 3, 20

Kaishev *et al.* **(2007)** V. K. Kaishev, D. S. Dimitrova e S. Haberman. Modelling the joint distribution of competing risks survival times using copula functions. *Insurance Mathematics and Economics*, 41: 339–361. Citado na pág. 21

Kalbfleisch e Prentice (1980) J. D. Kalbfleisch e R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Nova Iorque: John Wiley & Sons. Citado na pág. 1, 12, 19

Klein e Moeschberger (1988) J. P. Klein e M. L. Moeschberger. Bounds on net survival probabilities for dependent competing risks. *Biometrics*, 44:529–538. Citado na pág. 14

Klein e Moeschberger (2003) John P. Klein e Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Nova Iorque: Springer-Verlag, 2 ed. Citado na pág. 6, 66

Kuo e Yang (2000) L. Kuo e T. M. Yang. Bayesian reliability modeling for masked system lifetime. *Statistics and Probability Letters*, 47:229–241. Citado na pág. 19

Lin *et al.* **(2007)** T. I. Lin, J. C. Lee e S. Y. Yen. Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, 17:909–927. Citado na pág. 50, 51, 63, 66

Louzada-Neto (1999) F. Louzada-Neto. Polyhazard regression models for lifetime data. *Biometrics*, 55:1281–1285. Citado na pág. 19

Louzada-Neto *et al.* **(2002)** F. Louzada-Neto, J. J. Mazucheli e J. A. Achcar. Mixture hazard models for lifetime data. *Biometrics Journal*, 44:3–14. Citado na pág. 29

Louzada-Neto *et al.* **(2004)** F. Louzada-Neto, C. S. Andrade e F. R. Z. Almeida. On the non-identifiability problem arising on the poly-weibull model. *Communications in Statistics - Simulation and Computation*, 33(3):541–552. Citado na pág. 19

Maclean *et al.* **(1976)** C.J. Maclean, N. E. Morton, R.C. Elston e S. Yee. Skewness in commingled distributions. *Biometrics*, 32:695–699. Citado na pág. 50

Marín *et al.* **(2005)** J. M. Marín, M. T. Rodriguez-Bernal e M. P. Wiper. Using weibull mixture distributions to model heterogeneous survival data. *Communications in Statistics - Simulation and Computation*, 34:673–684. Citado na pág. 29

Mazucheli *et al.* **(2001)** J. Mazucheli, F. Louzada-Neto e J. A. Achcar. Bayesian inference for polyhazard models in the presence of covariates. *Computational Statistics & Data Analysis*, 38:1–14. Citado na pág. 19

Mazucheli *et al.* **(2012)** J. J. Mazucheli, F. Louzada-Neto e J. A. Achcar. The polysurvival model with long-term survivors. *Brazilian Journal of Probability and Statistics*, 26(3):313–324. Citado na pág. 33

Mclachlan e Peel (2000) G. Mclachlan e D. Peel. *Finite Mixture Models*. New York: John Wiley & Sons. Citado na pág. 15, 29, 48, 50

Nadarajah *et al.* **(2011)** S. Nadarajah, G. M. Cordeiro e E. M. M. Ortega. General results for the beta modified weibull distribution. *Journal of Statistical Computation and Simulation*, 81:1211–1232. Citado na pág. 26, 27

Nelsen (2006) R. B. Nelsen. *An Introduction to Copulas*. New York: Springer-Verlag, 2 ed. Citado na pág. 3, 20, 22

Oakes (1982) D. Oakes. A model for association in bivariate survival data. *Journal of the Royal Statistical Society, Series B*, 44:414–422. Citado na pág. 20

Peterson (1976) A. V. Peterson. Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks. *Proceedings of the National Academy of Sciences*, 73: 11–13. Citado na pág. 14

Priebe (1994) C. E. Priebe. Adaptive mixtures. *Journal of the American Statistical Association*, 89: 796–806. Citado na pág. 15

Richardson e Green (1997) S. Richardson e P. J. Green. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society B*, 59:731–792. Citado na pág. 48, 50

Romeo *et al.* **(2006)** J. S. Romeo, N. I. Tanaka e A. C. Pedroso de Lima. Bivariate survival modeling: a bayesian approach based on copulas. *Lifetime Data Analysis*, 12:205–222. Citado na pág. 20

Schwarz (1978) G. Schwarz. Estimating the dimensional of a model. *Annals of Statistics*, 6:461–464. Citado na pág. 45

Tanner e Wong (1987) M. A. Tanner e W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Society*, 82:528–550. Citado na pág. 66

Titterington *et al.* **(1985)** D. M. Titterington, A. F. M. Smith e U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley & Sons. Citado na pág. 15, 29

Trivedi e Zimmer (2005) P. K. Trivedi e D. M. Zimmer. Copula modelling: An introduction for practitioners. *Foundations and Trends in Econometrics*, 1:1–111. Citado na pág. 20, 22

Tsiatis (1975) A. Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences USA*, 72:20–22. Citado na pág. 25

Tsionas (2002) E. G. Tsionas. Bayesian analysis of finite mixtures of weibull distributions. *Communications in Statistics - Theory and Methods*, 31:37–48. Citado na pág. 29

Yashin *et al.* **(1986)** A.I. Yashin, K.G. Manton e E. Stallard. Dependent competing risks: a stochastic process model. *Journal of Mathematical Biology*, 24:119–140. Citado na pág. 22

Zheng e Klein (1994) M. Zheng e J. P. Klein. A self-consistent estimator of marginal survival functions based on dependent competing risk data and an assumed copula. *Communications in Statistics - Theory and Methods*, 23:2299–2311. Citado na pág. 14

Zheng e Klein (1995) M. Zheng e J. P. Klein. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82:127–138. Citado na pág. 25